

89748

Sixth Edition

METHODS

098

WILLIAM G. COCHRAN

*Professor of Statistics
Harvard University*



New Delhi

Bombay

Calcutta

GEORGE W. SNEDECOR is professor emeritus of statistics, Iowa State University, where he taught from 1913 to 1958 and where he was for fourteen years director of the statistical laboratory. His writings include a body of scientific journal articles, research bulletins, and books, including *Correlation and Machine Calculation* (with H. A. Wallace), *Calculation and Interpretation of Analysis of Variance and Covariance*, and *Statistical Methods*. He holds a master of science degree from the University of Michigan, and honorary doctor of science degrees from North Carolina State University and Iowa State University. He is a member of the International Statistical Institute, past president of the American Statistical Association, and an honorary Fellow of the British Royal Statistical Society. He has served also as consultant, Human Factors Division, U.S. Navy Electronics Laboratory, San Diego, California, where he now lives.

WILLIAM G. COCHRAN is professor of statistics, Harvard University. He has served formerly on the faculties of Johns Hopkins University, North Carolina State University, and Iowa State University. He holds master of arts degrees from Glasgow University and Cambridge University and an honorary master of arts degree from Harvard University. He is past president of the American Statistical Association, the Institute of Mathematical Statistics, and the Biometric Society. His writings include many research papers in the professional journals of his field; *Sampling Techniques*, 2nd ed., 1963; and *Experimental Designs* (with Gertrude M. Cox), 2nd ed., 1957.

© 1937, 1938, 1940, 1946, 1956, 1967 The Iowa State University Press
Ames, Iowa, U.S.A. All rights reserved

Sixth edition, 1967

Rs. 50/-

Indian Edition 1968 published by arrangement with the original American publishers The Iowa State University Press, U. S. A.

For Sale in India, Pakistan, Burma, Ceylon and Indonesia

Published by Mohan Pramlant for Oxford & IBH Publishing Co. Pvt. Ltd.,
60 Janpath, New Delhi 110 001 and printed at
Gopsons papers Pvt. Ltd., Noida

7-E7-11

Preface

In preparing the sixth edition we have kept in mind the two purposes this book has served during the past thirty years. Prior editions have been used extensively both as texts for introductory courses in statistics and as reference sources of statistical techniques helpful to research workers in the interpretation of their data.

As a text, the book contains ample material for a course extending throughout the academic year. For a one-term course, a suggested list of topics is given on the page preceding the Table of Contents. As in past editions, the mathematical level required involves little more than elementary algebra. Dependence on mathematical symbols has been kept to a minimum. We realize, however, that it is hard for the reader to use a formula with full confidence until he has been given proof of the formula or its derivation. Consequently, we have tried to help the reader's understanding of important formulas either by giving an algebraic proof where this is feasible or by explaining on common-sense grounds the roles played by different parts of the formula.

This edition retains also one of the characteristic features of the book—the extensive use of experimental sampling to familiarize the reader with the basic sampling distributions that underlie modern statistical practice. Indeed, with the advent of electronic computers, experimental sampling in its own right has become much more widely recognized as a research weapon for solving problems beyond the current skills of the mathematician.

Some changes have been made in the structure of the chapters, mainly at the suggestion of teachers who have used the book as a text. The former chapter 8 (Large Sample Methods) has disappeared, the retained material being placed in earlier chapters. The new chapter 8 opens with an introduction to probability, followed by the binomial and Poisson distributions (formerly in chapter 16). The discussion of multiple regression (chapter 13) now precedes that of covariance and multiple covariance (chapter 14).

Chapter 16 contains two related topics, the analysis of two-way classifications with unequal numbers of observations in the sub-classes and the analysis of proportions in two-way classifications. The first of these topics was formerly at the end of a long chapter on factorial arrangements; the second topic is new in this edition. This change seemed advisable for two reasons. During the past twenty years there has been a marked increase in observational studies in the social sciences, in medicine and public health, and in operations research. In their analyses, these studies often involve the handling of multiple classifications which present complexities appropriate to the later sections of the book.

Finally, in response to almost unanimous requests, the statistical tables in the book have been placed in an Appendix.

A number of topics appear for the first time in this edition. As in past editions, the selection of topics was based on our judgment as to those likely to be most useful. In addition to the new material on the analysis of proportions in chapter 16, other new topics are as follows:

- The analysis of data recorded in scales having only a small number of distinct values (section 5.8);
- In linear regression, the prediction of the independent variable X from the dependent variable Y , sometimes called linear calibration (section 6.14);
- Linear regression when X is subject to error (section 6.17);
- The comparison of two correlated estimates of variance (section 7.12);
- An introduction to probability (section 8.2);
- The analysis of proportions in ordered classifications (section 9.10);
- Testing a linear trend in proportions (section 9.11);
- The analysis of a set of 2×2 contingency tables (section 9.14);
- More extensive discussion of the effects of failures in the assumptions of the analysis of variance and of remedial measures (sections 11.10–11.13);
- Recent work on the selection of variates for prediction in multiple regression (section 13.13);
- The discriminant function (sections 13.14, 13.15);
- The general method of fitting non-linear regression equations and its application to asymptotic regression (sections 15.7–15.8).

Where considerations of space permitted only a brief introduction to the topic, references were given to more complete accounts.

Most of the numerical illustrations continue to be from biological investigations. In adding new material, both in the text and in the examples to be worked by the student, we have made efforts to broaden the

range of fields represented by data. One of the most exhilarating features of statistical techniques is the extent to which they are found to apply in widely different fields of investigation.

High-speed electronic computers are rapidly becoming available as a routine resource in centers in which a substantial amount of data are analyzed. Flexible standard programs remove the drudgery of computation. They give the investigator vastly increased power to fit a variety of mathematical models to his data; to look at the data from different points of view; and to obtain many subsidiary results that aid the interpretation. In several universities their use in the teaching of introductory courses in statistics is being tried, and this use is sure to increase.

We believe, however, that in the future it will be just as necessary that the investigator learn the standard techniques of analysis and understand their meaning as it was in the desk machine age. In one respect, computers may change the relation of the investigator to his data in an unfortunate way. When calculations are handed to a programmer who translates them into the language understood by the computer, the investigator, on seeing the printed results, may lack the self-assurance to query or detect errors that arose because the programmer did not fully understand what was wanted or because the program had not been correctly debugged. When data are being programmed it is often wise to include a similar example from this or another standard book as a check that the desired calculations are being done correctly.

For their generous permission to reprint tables we are indebted to the late Sir Ronald Fisher and his publishers, Oliver and Boyd; to Maxine Merrington, Catherine M. Thompson, Joyce N. May, E. Lord, and E. S. Pearson, whose work was published in *Biometrika*; to C. I. Bliss, E. L. Crow, C. White, and the late F. Wilcoxon; and to Bernard Ostle and his publishers, The Iowa State University Press. Thanks are due also to the many investigators who made data available to us as illustrative examples, and to teachers who gave helpful advice arising from their experience in using prior editions as a text. The work of preparing this edition was greatly assisted by a contract between the Office of Naval Research, Navy Department, and the Department of Statistics, Harvard University. Finally, we wish to thank Marianne Blackwell, Nancy Larson, James DeGracie and Richard Mensing for typing or proofreading, and especially Holly Lasewicz for her help at many stages of the work, including the preparation of the Indexes.

George W. Snedecor
William G. Cochran

A SHORT COURSE IN THE ELEMENTS OF STATISTICAL METHOD

CHAPTER	PAGES
1 Attributes.....	3- 31
2 Measurements.....	32- 61
3 Sampling distributions.....	{ 66- 74
	{ 77- 79
4 Comparison of two samples.....	91-104
5 Non-Parametric Methods.....	120-128
6 Regression.....	{ 135-145
	{ 149-157
7 Correlation.....	172-177
8 Binomial distribution.....	199-219
9 One-way classifications—Attributes.....	{ 228-231
	{ 236-238
10 One-way classifications—Measurements.....	258-271
11 Two-way classifications.....	299-311

Table of contents

Chapter 1. Sampling of Attributes

1.1	Introduction.....	3
1.2	Purpose of this chapter.....	4
1.3	The twin problems of sampling.....	4
1.4	A sample of farm facts. Point and interval estimates.....	5
1.5	Random sampling.....	10
1.6	Tables of random digits.....	12
1.7	Confidence interval: verification of theory.....	14
1.8	The sampled population.....	15
1.9	The frequency distribution and its graphical representation.....	16
1.10	Hypotheses about populations.....	20
1.11	Chi-square, an index of dispersion.....	20
1.12	The formula for chi-square.....	21
1.13	An experiment in sampling chi-square; the sampling distribution.....	22
1.14	Comparison with the theoretical distribution.....	25
1.15	The test of a null hypothesis or test of significance.....	26
1.16	Tests of significance in practice.....	28
1.17	Summary of technical terms.....	29

Chapter 2. Sampling From a Normally Distributed Population

2.1	Normally distributed population.....	32
2.2	Reasons for the use of the normal distribution.....	35
2.3	Tables of the normal distribution.....	35
2.4	Estimators of μ and σ	39
2.5	The array and its graphical representation.....	40
2.6	Algebraic notation.....	41
2.7	Deviations from sample mean.....	42
2.8	Another estimator of σ ; the sample standard deviation.....	44
2.9	Comparison of the two estimators of σ	46
2.10	Hints on the computation of s	47
2.11	The standard deviation of sample means.....	49
2.12	The frequency distribution of sample means.....	51
2.13	Confidence intervals for μ when σ is known.....	56
2.14	Size of sample.....	58
2.15	"Student's" t -distribution.....	59
2.16	Confidence limits for μ based on the t -distribution.....	61
2.17	Relative variation. Coefficient of variation.....	62

x Contents

Chapter 3. Experimental Sampling From a Normal Population

3.1	Introduction	66
3.2	A finite population simulating the normal	66
3.3	Random samples from a normal distribution	69
3.4	The distribution of sample means	70
3.5	Sampling distribution of s^2 and s	72
3.6	Interval estimates of σ^2	74
3.7	Test of a null hypothesis value of σ^2	76
3.8	The distribution of t	77
3.9	The interval estimate of μ : the confidence interval	78
3.10	Use of frequency distributions for computing \bar{X} and s	80
3.11	Computation of \bar{X} and s in large samples: example	81
3.12	Tests of normality	84
3.13	A test of skewness	86
3.14	Tests for kurtosis	86
3.15	Effects of skewness and kurtosis	88

Chapter 4. The Comparison of Two Samples

4.1	Estimates and tests of differences	91
4.2	A simulated paired experiment	92
4.3	Example of a paired experiment	94
4.4	Conditions for pairing	97
4.5	Tests of other null hypotheses about μ	97
4.6	Comparison of the means of two independent samples	100
4.7	The variance of a difference	100
4.8	A pooled estimate of variance	101
4.9	An experiment comparing two groups of equal size	102
4.10	Groups of unequal sizes	104
4.11	Paired versus independent groups	106
4.12	Precautions against bias-randomization	109
4.13	Sample size in comparative experiments	111
4.14	Analysis of independent samples when $\sigma_1 \neq \sigma_2$	114
4.15	A test of the equality of two variances	116

Chapter 5. Shortcut and Non-parametric Methods

5.1	Introduction	120
5.2	The t -test based on range	120
5.3	Median, percentiles and order statistics	123
5.4	The sign test	125
5.5	Non-parametric methods: ranking of differences between measurements	128
5.6	Non-parametric methods: ranking for unpaired measurements	130
5.7	Comparison of rank and normal tests	132
5.8	Scales with limited values	132

Chapter 6. Regression

6.1	Introduction	135
6.2	The regression of blood pressure on age	135
6.3	Shortcut methods of computation in regression	139
6.4	The mathematical model in linear regression	141
6.5	\hat{Y} as an estimator of $a + bX$	144
6.6	The estimator of σ_y	145

6.7	The method of least squares	147
6.8	The value of b in some simple cases.	147
6.9	The situation when X varies from sample to sample.	149
6.10	Interval estimates of β and tests of null hypotheses.	153
6.11	Prediction of the population regression line.	153
6.12	Prediction of an individual Y	155
6.13	Testing a deviation that looks suspiciously large.	157
6.14	Prediction of X from Y . Linear calibration.	159
6.15	Partitioning the sum of squares of the dependent variate.	160
6.16	Galton's use of the term "regression".	164
6.17	Regression when X is subject to error.	164
6.18	Fitting a straight line through the origin.	166
6.19	The estimation of ratios.	170
6.20	Summary.	170

Chapter 7. Correlation

7.1	Introduction.	172
7.2	The sample correlation coefficient r	173
7.3	Relation between the sample coefficients of correlation and regression.	175
7.4	The bivariate normal distribution.	177
7.5	Sampling variation of the correlation coefficient. Common elements.	181
7.6	Testing the null hypothesis $\rho = 0$	184
7.7	Confidence limits and tests of hypotheses about ρ	185
7.8	Practical utility of correlation and regression.	188
7.9	Variances of sums and differences of correlated variables.	190
7.10	The calculation of r in a large sample.	191
7.11	Non-parametric methods. Rank correlation.	193
7.12	The comparison of two correlated variances.	195

Chapter 8. Sampling From the Binomial Distribution

8.1	Introduction.	199
8.2	Some simple rules of probability.	199
8.3	The binomial distribution.	202
8.4	Sampling the binomial distribution.	205
8.5	Mean and standard deviation of the binomial distribution.	207
8.6	The normal approximation and the correction for continuity.	209
8.7	Confidence limits for a proportion.	210
8.8	Test of significance of a binomial proportion.	211
8.9	The comparison of proportions in paired samples.	213
8.10	Comparison of proportions in two independent samples: the 2×2 table.	215
8.11	Test of the independence of two attributes.	219
8.12	A test by means of the normal deviate z	220
8.13	Sample size for comparing two proportions.	221
8.14	The Poisson distribution.	223

Chapter 9. Attribute Data With More Than One Degree of Freedom

9.1	Introduction.	228
9.2	Single classifications with more than two classes.	228
9.3	Single classifications with equal expectations.	231
9.4	Additional tests.	233
9.5	The χ^2 test when the expectations are small.	235
9.6	Single classifications with estimated expectations.	236
9.7	Two-way classifications. The $2 \times C$ contingency table.	238

9.8	The variance test for homogeneity of the binomial distribution.....	240
9.9	Further examination of the data.....	242
9.10	Ordered classifications.....	243
9.11	Test for a linear trend in proportions.....	246
9.12	Heterogeneity χ^2 in testing Mendelian ratios.....	248
9.13	The $R \times C$ table.....	250
9.14	Sets of 2×2 tables.....	253

Chapter 10. One-Way Classifications. Analysis of Variance

10.1	Extension from two samples to many.....	258
10.2	An experiment with four samples.....	258
10.3	The analysis of variance.....	260
10.4	Effect of differences between the population means.....	264
10.5	The variance ratio, F	265
10.6	Analysis of variance with only two classes.....	267
10.7	Comparisons among class means.....	268
10.8	Inspection of all differences between pairs of means.....	271
10.9	Shortcut computation using ranges.....	275
10.10	Model I. Fixed treatment effects.....	275
10.11	Effects of errors in the assumptions.....	276
10.12	Samples of unequal sizes.....	277
10.13	Model II. Random effects.....	279
10.14	Structure of model II illustrated by sampling.....	282
10.15	Confidence limits for σ_A^2	284
10.16	Samples within samples. Nested classifications.....	285
10.17	Samples within samples. Mixed model.....	288
10.18	Samples of unequal sizes. Random effects.....	289
10.19	Samples within samples. Unequal sizes.....	291
10.20	Intraclass correlation.....	294
10.21	Tests of homogeneity of variance.....	296

Chapter 11. Two-Way Classifications

11.1	Introduction.....	299
11.2	An experiment with two criteria of classification.....	299
11.3	Comparisons among means.....	301
11.4	Algebraic notation.....	302
11.5	Mathematical model for a two-way classification.....	303
11.6	Partitioning the treatments sum of squares.....	308
11.7	Efficiency of blocking.....	311
11.8	Latin squares.....	312
11.9	Missing data.....	317
11.10	Non-conformity to model.....	321
11.11	Gross errors: rejection of extreme observations.....	321
11.12	Lack of independence in the errors.....	323
11.13	Unequal error variances due to treatments.....	324
11.14	Non-normality. Variance-stabilizing transformations.....	325
11.15	Square-root transformation for counts.....	325
11.16	Arcsin transform for proportions.....	327
11.17	The logarithmic transformation.....	329
11.18	Non-additivity.....	330
11.19	Tukey's test of additivity.....	331
11.20	Non-additivity in a Latin square.....	334

Chapter 12. Factorial Experiments

12.1	Introduction	339
12.2	The single factor versus the factorial approach	339
12.3	Analysis of the 2^2 factorial experiment	342
12.4	The 2^2 factorial when interaction is present	344
12.5	The general two-factor experiment	346
12.6	Response curves	349
12.7	Response curves in two-factor experiments	352
12.8	Example of a response surface	354
12.9	Three-factor experiments; the 2^3	359
12.10	Three-factor experiments; a $2 \times 3 \times 4$	361
12.11	Expected values of mean squares	364
12.12	The split-plot or nested design	369
12.13	Series of experiments	375
12.14	Experiments with perennial crops	377

Chapter 13. Multiple Regression

13.1	Introduction	381
13.2	Two independent variables	381
13.3	The deviations mean square and the F -test	385
13.4	Alternative method of calculation. The inverse matrix	389
13.5	Standard errors of estimates in multiple regression	391
13.6	The interpretation of regression coefficients	393
13.7	Relative importance of different X -variables	398
13.8	Partial and multiple correlation	400
13.9	Three or more independent variables. Computations	403
13.10	Numerical example. Computing the b 's	405
13.11	Numerical example. Computing the inverse matrix	409
13.12	Deletion of an independent variable	412
13.13	Selection of variates for prediction	412
13.14	The discriminant function	414
13.15	Numerical example of the discriminant function	416

Chapter 14. Analysis of Covariance

14.1	Introduction	419
14.2	Covariance in a completely randomized experiment	420
14.3	The F -test of the adjusted means	424
14.4	Covariance in a 2-way classification	425
14.5	Interpretation of adjusted means in covariance	429
14.6	Comparison of regression lines	432
14.7	Comparison of the "Between Classes" and the "Within Classes" regressions	436
14.8	Multiple covariance	438
14.9	Multiple covariance in a 2-way table	443

Chapter 15. Curvilinear Regression

15.1	Introduction	447
15.2	The exponential growth curve	449
15.3	The second degree polynomial	453
15.4	Data having several Y 's at each X value	456
15.5	Test of departure from linear regression in covariance analysis	460

15.6	Orthogonal polynomials.....	460
15.7	A general method of fitting non-linear regressions.....	465
15.8	Fitting an asymptotic regression.....	467

Chapter 16. Two-Way Classifications With Unequal Numbers and Proportions

16.1	Introduction.....	472
16.2	Unweighted analysis of cell means.....	475
16.3	Equal numbers within rows.....	477
16.4	Proportional sub-class numbers.....	478
16.5	Disproportionate numbers. The 2×2 table.....	483
16.6	Disproportionate numbers. The $R \times 2$ table.....	484
16.7	The $R \times C$ table. Least squares analysis.....	488
16.8	The analysis of proportions in 2-way tables.....	493
16.9	Analysis in the p scale: a 2×2 table.....	495
16.10	Analysis in the p scale: a 3×2 table.....	496
16.11	Analysis of logits in an $R \times C$ table.....	497
16.12	Numerical example.....	498

Chapter 17. Design and Analysis of Sampling

17.1	Populations.....	504
17.2	A simple example.....	505
17.3	Probability sampling.....	508
17.4	Listing the population.....	509
17.5	Simple random sampling.....	511
17.6	Size of sample.....	516
17.7	Systematic sampling.....	519
17.8	Stratified sampling.....	520
17.9	Choice of sample sizes in the individual strata.....	523
17.10	Stratified sampling for attributes.....	526
17.11	Sampling in two stages.....	528
17.12	The allocation of resources in two-stage sampling.....	531
17.13	Selection with probability proportional to size.....	534
17.14	Ratio and regression estimates.....	536
17.15	Further reading.....	538

Appendix

List of Appendix Tables and Notes.....	541
Appendix Tables.....	543
Author Index.....	577
Index to Numerical Examples.....	581
Subject Index.....	585

STATISTICAL METHODS



Sampling of attributes

1.1—Introduction. The subject matter of the field of statistics has been described in various ways. According to one definition, statistics deals with techniques for collecting, analyzing, and drawing conclusions from data. This description helps to explain why an introduction to statistical methods is useful to students who are preparing themselves for a career in one of the sciences and to persons working in any branch of knowledge in which much quantitative research is carried out. Such research is largely concerned with gathering and summarizing observations or measurements made by planned experiments, by questionnaire surveys, by the records of a sample of cases of a particular kind, or by combing past published work on some problem. From these summaries, the investigator draws conclusions that he hopes will have broad validity.

The same intellectual activity is involved in much other work of importance. Samples are extensively used in keeping a continuous watch on the output of production lines in industry, in obtaining national and regional estimates of crop yields and of business and employment conditions, in the auditing of financial statements, in checking for the possible adulteration of foods, in gauging public opinion and voter preferences; in learning how well the public is informed on current issues, and so on.

Acquaintance with the main ideas in statistical methodology is also an appropriate part of a general education. In newspapers, books, television, radio, and speeches we are all continuously exposed to statements that draw general conclusions: for instance, that the cost of living rose by 0.3% in the last month, that the smoking of cigarettes is injurious to health, that users of "Blank's" toothpaste have 23% fewer cavities, that a television program had 18.6 million viewers. When an inference of this kind is of interest to us, it is helpful to be able to form our own judgment about the truth of the statement. Statistics has no magic formula for doing this in all situations, for much remains to be learned about the problem of making sound inferences. But the basic ideas in statistics assist us in thinking clearly about the problem, provide some guidance about the conditions that must be satisfied if sound inferences are to be made, and enable us to detect many inferences that have no good logical foundation.

1.2—Purpose of this chapter. Since statistics deals with the collection, analysis, and interpretation of data, a book on the subject might be expected to open with a discussion of methods for collecting data. Instead, we shall begin with a simple and common type of data already collected, the replies to a question given by a sample of the farmers in a county, and discuss the problem of making a statement from this sample that will apply to all farmers in the county. We begin with this problem of making inferences beyond the data because the type of inference that we are trying to make governs the way in which the data must be collected. In earlier days, and to some extent today also, many workers did not appreciate this fact. It was a common experience for statisticians to be approached with: Here are my results. What do they show? Too often the data were incapable of showing anything that would have been of interest to an investigator, because the method of collecting the data failed to meet the conditions needed for making reliable inferences beyond the data.

In this chapter, some of the principal tools used in statistics for making inferences will be presented by means of simple illustrations. The mathematical basis of these tools, which lies in the theory of probability, will not be discussed until later. Consequently, do not expect to obtain a full understanding of the techniques at this stage, and do not worry if the ideas seem at first unfamiliar. Later chapters will give you further study of the properties of these techniques and enhance your skill in applying them to a broad range of problems.

1.3—The twin problems of sampling. A *sample* consists of a small collection from some larger aggregate about which we wish information. The sample is examined and the facts about it learned. Based on these facts, the problem is to make correct inferences about the *aggregate* or *population*. It is the sample that we observe, but it is the population which we seek to know.

This would be no problem were it not for ever-present variation. If all individuals were alike, a sample consisting of a single one would give complete information about the population. Fortunately, there is endless variety among individuals as well as their environments. A consequence is that successive samples are usually different. Clearly, the facts observed in a sample cannot be taken as facts about the population. Our job then is to reach appropriate conclusions about the population despite sampling variation.

But not every sample contains information about the population sampled. Suppose the objective of an experimental sampling is to determine the growth rate in a population of young mice fed a new diet. Ten of the animals are put in a cage for the experiment. But the cage gets located in a cold draught or in a dark corner. Or an unnoticed infection spreads among the mice in the cage. If such things happen, the growth rate in the sample may give no worthwhile information about that in the population of normal mice. Again, suppose an interviewer in an opinion

poll picks only families among his friends whom he thinks it will be pleasant to visit. His sample may not at all represent the opinions of the population. This brings us to a second problem: to collect the sample in such a way that the sought-for information is contained in it.

So we are confronted with the twin problems of the investigator: to design and conduct his sampling so that it shall be representative of the population; then, having studied the sample, to make correct inferences about the sampled population.

1.4—A sample of farm facts. Point and interval estimates. In 1950 the USDA Division of Cereal and Forage Insect Investigations, cooperating with the Iowa Agricultural Experiment Station, conducted an extensive sampling in Boone County, Iowa, to learn about the interrelation of factors affecting control of the European corn borer.* One objective of the project was to determine the extent of spraying or dusting for control of the insects. To this end a random sample of 100 farmers were interviewed; 23 of them said they applied the treatment to their corn fields. Such are the facts of the sample.

What *inferences* can be made about the population of 2,300 Boone County farmers? There are two of them. The first is described as a *point estimate*, while the second is called an *interval estimate*.

1. The *point estimate* of the fraction of farmers who sprayed is 23%, the same as the sample ratio; that is, an estimated 23% of Boone County farmers sprayed their corn fields in 1950. This may be looked upon as an average of the numbers of farmers per hundred who sprayed. From the actual count of sprayers in a single hundred farmers it is inferred that the average number of sprayers in all possible samples of 100 is 23.

This sample-to-population inference is usually taken for granted. Most people pass without a thought from the sample fact to this inference about the population. Logically, the two concepts are distinct. It is wise to examine the procedure of the sampling before attributing to the population the percentage reported in a sample.

2. An *interval estimate* of the point is made by use of table 1.4.1. In the first part of the table, indicated by 95% in the heading, look across the top line to the sample size of 100, then down the left-hand column to the number (or frequency) observed, 23 farmers. At the intersection of the column and line you will find the figures 15 and 32. The meaning is this: one may be confident that the true percentage in the sampled population lies in the interval from 15% to 32%. This interval estimate is called the *confidence interval*. The nature of our confidence will be explained later.

In summary: based on a random sample, we said first that our estimate of the percentage of sprayers in Boone County was 23%, but we gave no indication of the amount by which the estimate might be in error. Next we asserted confidently that the true percentage was not farther from our point estimate, 23%, than 8 percentage points below or 9 above.

Let us illustrate these concepts in another fashion. Imagine a bin

* Data furnished courtesy of Dr. T. A. Brindley

TABLE 1 4 1
95% CONFIDENCE INTERVAL (PER CENT) FOR BINOMIAL DISTRIBUTION (1)*

Number Observed f	Size of Sample n						Fraction Observed f/n	Size of Sample		
	10	15	20	30	50	100		250	1000	
0	0 27	0 20	0 15	0 10	0 07	0 4	0 00	0 1	0 0	0
1	0 40	0 31	0 23	0 17	0 11	0 5	01	0 4	0 2	2
2	3 61	2 37	1 30	1 21	0 14	0 7	02	1 5	1 3	3
3	8 62	5 45	4 36	2 25	1 17	1 8	03	1 6	2 4	5
4	15 74	9 56	7 42	4 30	2 19	1 10	04	2 7	3 5	7
5	22 78	14 64	10 47	6 33	3 22	2 11	05	3 9	4 7	9
6	26 85	19 67	14 54	9 37	5 24	2 12	06	3 10	5 8	11
7	38 92	19 71	14 59	10 41	6 27	3 14	07	4 11	6 9	13
8	39 97	29 81	20 65	13 44	7 29	4 15	08	5 12	6 10	15
9	60 100	33 81	22 71	16 48	9 31	4 16	09	6 13	7 11	17
10	73 100	36 86	29 71	17 53	10 34	5 18	10	7 14	8 12	19
11		44 91	29 78	20 56	12 36	5 19	11	7 16	9 13	21
12		55 95	35 80	23 60	13 38	6 20	12	8 17	10 14	23
13		63 98	41 86	24 64	15 41	7 21	13	9 18	11 15	25
14		69 100	47 86	29 68	16 43	8 22	14	10 19	12 16	27
15		80 100	53 90	32 68	18 44	9 24	15	10 20	13 17	29
16			58 93	32 71	20 46	9 25	16	11 21	14 18	31
17			64 96	36 76	21 48	10 26	17	12 22	15 19	33
18			70 99	40 77	23 50	11 27	18	13 23	16 21	35
19			77 100	44 80	25 53	12 28	19	14 24	17 22	37
20			85 100	47 83	27 55	13 29	20	15 26	18 23	39
21				52 84	28 57	14 30	21	16 27	19 24	41
22				56 87	30 59	14 31	22	17 28	19 25	43
23				59 90	32 61	15 32	23	18 29	20 26	45
24				63 91	34 63	16 33	24	19 30	21 27	47
25				67 94	36 64	17 35	25	20 31	22 28	49
26				70 96	37 66	18 36	26	20 32	23 29	51
27				75 98	39 68	19 37	27	21 33	24 30	53
28				79 99	41 70	19 38	28	22 34	25 31	55
29				83 100	43 72	20 39	29	23 35	26 32	57
30				90 100	45 73	21 40	30	24 36	27 33	59
31					47 75	22 41	31	25 37	28 34	61
32					50 77	23 42	32	26 38	29 35	63
33					52 79	24 43	33	27 39	30 36	65
34					54 80	25 44	34	28 40	31 37	67
35					56 82	26 45	35	29 41	32 38	69
36					57 84	27 46	36	30 42	33 39	71
37					59 85	28 47	37	31 43	34 40	73
38					62 87	28 48	38	32 44	35 41	75
39					64 88	29 49	39	33 45	36 42	77
40					66 90	30 50	40	34 46	37 43	79
41					69 91	31 51	41	35 47	38 44	81
42					71 93	32 52	42	36 48	39 45	83
43					73 94	33 53	43	37 49	40 46	85
44					76 95	34 54	44	38 50	41 47	87
45					78 97	35 55	45	39 51	42 48	89
46					81 98	36 56	46	40 52	43 49	91
47					83 99	37 57	47	41 53	44 50	93
48					86 100	38 58	48	42 54	45 51	95
49					89 100	39 59	49	43 55	46 52	97
50					93 100	40 60	50	44 56	47 53	99
					†			††	††	

* Reference (1) at end of chapter

† If f exceeds 50 read $100 - f =$ number observed and subtract each confidence limit from 100†† If f/n exceeds 0.50 read $100 - f/n =$ fraction observed and subtract each confidence limit from 100

TABLE 141 --(Continued)
99% CONFIDENCE INTERVAL (PER CENT) FOR BINOMIAL DISTRIBUTION (1)*

Number Observed f	Size of Sample n										Fraction Observed f/n	Size of Sample	
	10	15	20	30	50	100						250	1000
0	0 38	0 28	0 21	0 16	0 10	0 5	0 00	0	2	0	1		
1	0 52	0 38	0 30	0 21	0 14	0 7	01	0	5	0	2		
2	1 63	1 47	0 38	0 26	0 17	0 9	02	1	6	1	3		
3	4 71	3 54	2 43	1 31	1 20	0 10	03	1	7	2	4		
4	9 79	5 63	4 50	2 35	1 23	1 12	04	2	9	3	6		
5	15 85	9 68	6 58	4 39	2 26	1 13	05	2	10	3	7		
6	21 91	13 73	9 61	6 43	3 29	2 14	06	3	11	4	8		
7	29 96	17 78	12 64	8 47	4 31	2 16	07	3	13	5	9		
8	37 99	22 83	16 71	10 51	6 33	3 17	08	4	14	6	10		
9	48 100	27 87	20 73	12 54	7 36	3 18	09	5	15	7	12		
10	62 100	32 91	20 80	15 57	8 38	4 19	10	6	16	8	13		
11		37 95	27 80	15 62	10 40	4 20	11	6	17	9	14		
12		46 97	29 84	19 66	11 43	5 21	12	7	18	9	15		
13		53 99	36 88	20 68	12 45	6 23	13	8	19	10	16		
14		62 100	39 91	24 70	14 47	6 24	14	9	20	11	17		
15		72 100	42 94	25 75	15 49	7 26	15	9	22	12	18		
16			50 96	30 76	17 51	8 27	16	10	23	13	19		
17			57 98	32 80	18 53	9 29	17	11	24	14	20		
18			62 100	34 81	20 55	9 30	18	12	25	15	21		
19			70 100	38 85	21 57	10 31	19	13	26	16	22		
20			79 100	43 85	23 59	11 32	20	14	27	17	23		
21				46 88	24 61	12 33	21	15	28	18	24		
22				49 90	26 63	12 34	22	16	30	19	26		
23				53 92	28 65	13 35	23	17	31	20	27		
24				57 94	29 67	14 36	24	18	32	21	28		
25				61 96	31 69	15 38	25	18	33	22	29		
26				65 98	33 71	16 39	26	19	34	22	30		
27				69 99	35 72	16 40	27	20	35	23	31		
28				74 100	37 74	17 41	28	21	36	24	32		
29				79 100	39 76	18 42	29	22	37	25	33		
30				84 100	41 77	19 43	30	23	38	26	34		
31					43 79	20 44	31	24	39	27	35		
32					45 80	21 45	32	25	40	28	36		
33					47 82	21 46	33	26	41	29	37		
34					49 83	22 47	34	26	42	30	38		
35					51 85	23 48	35	27	43	31	39		
36					53 85	24 49	36	28	44	32	40		
37					55 88	25 50	37	29	45	33	41		
38					57 89	26 51	38	30	46	34	42		
39					60 90	27 52	39	31	47	35	43		
40					62 92	28 53	40	32	48	36	44		
41					64 93	29 54	41	33	50	37	45		
42					67 94	29 55	42	34	51	38	46		
43					69 96	30 56	43	35	52	39	47		
44					71 97	31 57	44	36	53	40	48		
45					74 93	32 58	45	37	54	41	49		
46					77 99	33 59	46	38	55	42	50		
47					80 99	34 60	47	39	55	43	51		
48					83 100	35 61	48	40	56	44	52		
49					86 100	36 62	49	41	57	45	53		
50					90 100	37 63	50	42	58	46	54		

+

++

++

* Reference (1) at end of chapter

† If f exceeds 50 read $100 - f$ = number observed and subtract each confidence limit from 100

†† If f/n exceeds 0.50 read $1.00 - f/n$ = fraction observed and subtract each confidence limit from 100

8 Chapter 1: Sampling of Attributes

filled with beans, some white and some colored, thoroughly mixed. Dip out a scoopful of them at random, count the number of each color and calculate the percentage of white, say 40%. Now this is not only a count of the percentage of white beans in the sample but it is an estimate of the fraction of white beans in the bin. How close an estimate is it? That is where the second inference comes in. If there were 250 beans in the scoop, we look at the table for size of sample 250, fraction observed = 0.40. From the table we say with confidence that the percentage of white beans in the bin is between 34% and 46%.

So far we have given no measure of the amount of confidence which can be placed in the second inference. The table heading is "95% Confidence Interval," indicating a degree of confidence that can be described as follows: If the sampling is repeated indefinitely, each sample leading to a new confidence interval (that is, to a new interval estimate), then in 95% of the samples the interval will cover the true population percentage. If one makes a practice of sampling and if for each sample he states that the population percentage lies within the corresponding confidence interval, about 95% of his statements will be correct. Other and briefer descriptions will be proposed later.

If you feel unsafe in making inferences with the chance of being wrong in 5% of your statements, you may use the second part of the table, "99% Confidence Interval." For the Boone County sampling the interval widens to 13%–35%. If one says that the population percentage lies within these limits, he will be right unless a one-in-a-hundred chance has occurred in the sampling.

If the size of the population is known, as it is in the case of Boone County farmers, the point and interval estimates can be *expanded* from percentages to numbers of individuals. There were 2,300 farmers in the county. Thus we estimate the number of sprayers in Boone County in 1950 as

$$(0.23)(2,300) = 529 \text{ farmers}$$

In the same way, since the 95% confidence interval extends from 15% to 32% of the farmers, the 95% limits for the number of farmers who sprayed are

$$(0.15)(2,300) = 345 \text{ farmers: and } (0.32)(2,300) = 736 \text{ farmers}$$

Two points about interval estimates need emphasis. First, the confidence statement is a statement about the population ratio, *not about the ratio in other samples that might be drawn*. Second, the uncertainty involved comes from the sampling process. Each sample specifies an interval estimate. Whether or not the interval happens to include the fixed population ratio is a hazard of the process. Theoretically, the 95% confidence intervals are determined so that 95% of them will cover the true value.

Before a sample is drawn, one can specify the probability of the truth

of his prospective confidence statement. He can say, "I expect to take a random sample and to make an interval estimate from it. The probability is 0.95 that the interval will cover the population fraction." After the sample is drawn, however, the confidence statement is either true or it is false. Consequently, in reporting the results of the Boone County sampling, it would be incorrect to say, "The probability is 0.95 that the number of sprayers in Boone County in 1950 lies between 345 and 736." This logical point is a subtle one, and does not weaken the effectiveness of confidence interval statements. In a specific application, we do not know whether our confidence statement is one of the 95% that are correct or one of the 5% that are wrong. There are methods, in particular the method known as the Bayesian approach, that provide more definite probability statements about a single specific application, but they require more assumptions about the nature of the population that is being sampled.

The heading of this chapter is "Sampling of Attributes." In the numerical example the attribute in question was whether the farm had been sprayed or not. The possession or lack of an attribute distinguishes the two classes of individuals making up the population. The data from the sample consist of the numbers of members of the sample found to have or to lack the attribute under investigation. The sampling of populations with two attributes is very common. Examples are *Yes* or *No* answers to a question, *Success* or *Failure* in some task, patients *Improved* or *Not Improved* under a medical treatment, and persons who *Like* or *Dislike* some proposal. Later (chapter 9) we shall study the sampling of populations that have more than two kinds of attributes, such as persons who are *Strongly Favorable*, *Mildly Favorable*, *Neutral*, *Mildly Unfavorable*, or *Strongly Unfavorable* to some proposal. The theory and methods for measurement data, such as heights, weights, or ages, will be considered in chapter 2.

This brief preview displays a goodly portion of the wares that the statistician has to offer: the sampling of populations, examination of the facts turned up by the sample, and, based on these facts, inferences about the sampled population. Before going further, you may clarify your thinking by working a few examples.

Examples form an essential part of our presentation of statistics. In each list they are graded so that you may start with the easier. It is suggested that a few in each group be worked after the first reading of the text, reserving the more difficult until experience is enlarged. Statistics cannot be mastered without this or similar practice.

EXAMPLE 1.4.1—In controlling the quality of a mass-produced article in industry, a random sample of 100 articles from a large lot were each tested for effectiveness. Ninety-two were found effective. What are the 99% confidence limits for the percentage of effective articles in the whole lot? Ans. 83% and 97%. Hint: look in the table for $100 - 92 = 8$

EXAMPLE 1.4.2—If 1,000 articles in the preceding example had been tested and only 8% found ineffective, what would be the 99% limits? Ans. Between 90% and 94% are effective. Note how the limits have narrowed as a result of the increased sample size.

10 Chapter 1: Sampling of Attributes

EXAMPLE 1.4.3—A sampler of public opinion asked 50 men to express their preferences between candidates A and B. Twenty preferred A. Assuming random sampling from a population of 5,000, the sampler stated that between 1,350 and 2,750 in the population preferred A. What confidence interval was he using? Ans. 95%.

EXAMPLE 1.4.4—In a health survey of adults, 86% stated that they had had measles at some time in the past. On the basis of this sample the statistician asserted that unless a 1-in-20 chance had occurred, the percentage of adults in the population who had had measles was between 81% and 90%. Assuming random sampling, what was the size of the sample? Ans. 250. Note: the statistician's inference may have been incorrect for other reasons. Some people have a mild attack of measles without realizing it. Others may have forgotten that they had it. Consequently, the confidence limits may be underestimates for the percentage in the population who actually had measles, as distinct from the percentage who would state that they had it.

EXAMPLE 1.4.5—If in the sample of 100 Boone County farmers none had sprayed, what 95% confidence statement would you make about the farmers in the county? Ans. Between none and 4% sprayed. But suppose that all farmers in the sample were sprayers, what is the 99% confidence interval? Ans. 95%–100%.

EXAMPLE 1.4.6—If you guess that in a certain population between 25% and 75% of the housewives own a specified appliance, and if you wish to draw a sample that will, at the 95% confidence level, yield an estimate differing by not more than 6 from the correct percentage, about how large a sample must you take? Ans. 250.

EXAMPLE 1.4.7—An investigator interviewed 115 women over 40 years of age from the lower middle economic level in rural areas of middlewestern states. Forty-six of them had listened to a certain radio program three or more times during the preceding month. Assuming random sampling, what statement can be made about the percentage of women listening in the population, using the 99% interval? Ans. Approximately, between 28.4% and 52.5% listen. You will need to interpolate between the results for $n = 100$ and $n = 250$. Appendix A 1 (p. 541) gives hints on interpolation.

EXAMPLE 1.4.8—For samples that show 50% in a certain class, write down the width of the 95% confidence interval for $n = 10, 20, 30, 50, 100, 250$, and 1,000. For each sample size n , multiply the width of the interval by \sqrt{n} . Show that the product is always near 200. This means that the width of the interval is approximately related to the sample size by the formula $W = 200/\sqrt{n}$. We say that the width goes down as $1/\sqrt{n}$.

1.5—Random sampling. The confidence intervals in table 1 4.1 were computed mathematically on the assumption that the data are a random sample from the population. In its simplest form, random sampling means that every member of the population has an equal chance of appearing in the sample, independently of the other members that happen to fall in the sample. Suppose that the population has four members, numbered 1, 2, 3, 4, and that we are drawing samples of size two. There are ten possible samples that contain two members: namely, (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4), (1, 1), (2, 2), (3, 3), and (4, 4). With simple random sampling, each of these ten samples has an equal chance of being the sample that is drawn. Notice two things. Every member appears once in three samples and twice in one sample, so that the sampling shows no favoritism as between one member and another. Secondly, look at the four samples in which a 1 appears, (1, 2), (1, 3), (1, 4), and (1, 1). The second member is equally likely to be a 1, 2, 3, or 4. Thus, if we are told that 1 has been drawn as the first member of the sample, we know that



each member of the population still has an equal chance of being the second member of the sample. This is what is meant by the phrase "independently of the other members that happen to fall in the sample."

A common variant of this method of sampling is to allow any member of the population to appear only *once* in the sample. There are then six possible samples of size two: (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), and (3, 4). This is the kind of sampling that occurs when two numbers are drawn out of a hat, no number being replaced in the hat. This type of sampling is called *random sampling without replacement*, whereas the sampling described in the preceding paragraph is *random sampling with replacement*. If the sample is a small fraction of the population, the two methods are practically identical, since the possibility that the same item appears more than once in a sample is negligible. Throughout most of the book we shall not distinguish between the two methods. In chapter 17, formulas applicable to sampling without replacement are presented.

There are more complex types of random sampling. In all of them, every member of the population has a *known* probability of coming into the sample, but these probabilities may not be equal or they may depend, in a known way, on the other members that are in the sample. In the Boone County sampling a book was available showing the location of every farm in the county. Each farm was numbered so that a random sample could have been drawn by mixing the numbers thoroughly in a box, then having a hundred of them drawn by a blindfolded person. Actually, the samplers used a scheme known as *stratified random sampling*. From the farms in each township (a subdivision of the county) they drew a random sample with a size proportional to the number of farms in that township. In this example, each farm still has an equal chance of appearing in the sample, but the sample is constructed to contain a specified number from every township. The chief advantage is to spread the sample more uniformly over the county, retaining the principle of randomness within each township. Statistical methods for stratified samples are presented in chapter 17. The conclusions are only slightly altered by considering the sample completely random. Unless otherwise mentioned, we will use the phrases "random sample" and "random sampling" to denote the simplest type of random sampling with replacement as described in the first paragraph of this section.

An important feature of all random sampling schemes is that the sampler has no control over the specific choice of the units that appear in the sample. If he exercises judgment in this selection, by choosing "typical" members or excluding members that appear "atypical," his results are not amenable to probability theory, and confidence intervals, which give valuable information about the accuracy of estimates made from the sample, cannot be constructed.

In some cases the population is thoroughly mixed before the sample is taken, as illustrated by the macerating and blending of food or other chemical products, by a naturally mixed aggregate such as the blood

stream, or by the sampling of a liquid from a vat that has been repeatedly stirred. Given an assurance of thorough mixing, the sample can be drawn from the most accessible part of the population, because any sample should give closely similar results. But complete mixing in this sense is often harder to achieve than is realized. With populations that are variable but show no clear pattern of variation, there is a temptation to conclude that the population is naturally mixed in a random fashion, so that any convenient sample will behave like one randomly drawn. This assumption is hazardous, and is difficult to verify without a special investigation.

One way of drawing a random sample is to list the members of the population in some order and write these numbers on slips of paper, marbles, beans, or small pieces of cardboard. These are placed in a box or bag, mixed carefully, and drawn out, with eyes shut, one by one until the desired size of sample is reached. With small populations this method is convenient, and was much used in the past for classroom exercises. It has two disadvantages. With large populations it is slow and unwieldy. Further, tests sometimes show that if a large number of samples are drawn, the samples differ from random samples in a noticeable way, for instance by having certain members of the population present more frequently than they should be. In other words, the mixing was imperfect.

1.6—Tables of random digits. Nowadays, samples are mostly drawn by the use of tables of random digits. These tables are produced by a process—usually mechanical or electrical—that gives each of the digits from 0 to 9 an equal chance of appearing at every draw. Before publication of the tables, the results of the drawings are checked in numerous ways to ensure that the tables do not depart materially from randomness in a manner that would vitiate the commonest usages of the tables. Table A 1 (p. 543) contains 10,000 such digits, arranged in 5×5 blocks to facilitate reading. There are 100 rows and 100 columns, each numbered from 00 to 99. Table 1.6.1 shows the first 100 numbers from this table.

The chaotic appearance of the set of numbers is evident. To illustrate how the table is used with attribute data, suppose that 50% of the members of a population answer "Yes" to some question. We wish to study how well the proportion answering "Yes" is estimated from a sam-

TABLE 1.6.1
ONE HUNDRED RANDOM DIGITS FROM TABLE A 1

	00-04	05-09	10-14	15-19
00	54463	22662	65905	70639
01	15389	85205	18850	39226
02	85941	40756	82414	02015
03	61149	69440	11286	88218
04	05219	51619	10651	67079

ple of size 20. A "Yes" answer can be represented by the appearance of one of the digits 0, 1, 2, 3, 4, or alternatively by the appearance of an *odd* digit. With either choice, the probability of a "Yes" at any draw in the table is one-half. We shall choose the digits 0, 1, 2, 3, 4 to represent "Yes," and let each row represent a different sample of size 20. A count, much quicker than drawing slips of paper from a box, shows that the successive rows in table 1.6.1 contain 9, 9, 12, 11, and 9 "Yes" answers. Thus, the proportions of "Yes" answers in these five samples of size 20 are, respectively, 0.45, 0.45, 0.60, 0.55, and 0.45. Continuing in this way we can produce estimates of the proportion of "Yes" answers given by a large number of separate samples of size 20, and then examine how close the estimates are to the population value. In counting the row numbered 02, you may notice a run of results that is typical of random sampling. The row ends with a succession of eight consecutive "Yes" answers, followed by a single "No." Observing this phenomenon by itself, one might be inclined to conclude that the proportion in the population must be larger than one-half, or that something is wrong with the sampling process.

Table A 1 can also be used to investigate sampling in which the proportion in the population is any of the numbers 0.1, 0.2, 0.3, . . . 0.9. With 0.3, for example, we let the digits 0, 1, or 2 represent the presence of the attribute and the remaining seven digits its absence. If you are interested in a population in which the proportion is 0.37, the method is to select pairs of digits, letting any pair between 00 and 36 denote the presence of the attribute. Tables of random digits are employed in studying a wide range of sampling problems. You can probably see how to use them to answer such questions as: On the average, how many digits must be taken until a 1 appears?—or, How frequently does a 3 appear before either a 1 or a 9 has appeared? In fact, sampling from tables of random digits has become an important technique for solving difficult problems in probability for which no mathematical solution is known at present. This technique goes by the not inappropriate name of the *Monte Carlo method*. For this reason, modern electronic computing machines have programs available for creating their own tables of random digits as they proceed with their calculations.

To the reader who is using random numbers for his own purposes, we suggest that he start on the first page and proceed systematically through the table. At the end of any problem, note the rows and columns used and the direction taken in counting. This is sometimes needed for later reference or in communicating the results to others. Since no digit is used more than once, the table may become exhausted, but numerous tables are available. Reference (2) contains 1 million digits. In classroom use, when a number of students are working from the same table, obtaining samples whose results will be put together, different students can start at different parts of the table and also vary the direction in which they proceed, in order to avoid duplicating the results of others.

1.7—Confidence interval: verification of theory. One who draws samples from a known population is likely to be surprised at the capricious way in which the items turn up. It is a salutary discipline for a student or investigator to observe the laws of chance in action lest he become too confident of his professional samplings. At this point we recommend that a number of samples be selected from a population in which the proportion of "Yes" answers is one-half. Vary the sample sizes, choosing some of each of the sizes 10, 15, 20, 30, 50, 100, and 250 for which confidence intervals are given in table 1.4.1 (1,000 is too large). For each sample, record the sample sizes and the set of rows and columns used in the table of random digits. From the number of "Yes" answers and the sample size, read table 1.4.1 to find the 95% and 99% confidence intervals for the percentage of "Yes" answers in the population. For each sample, you can then verify whether the confidence interval actually covers 50%. If possible, draw 100 or more samples, since a large number of samples is necessary for any close verification of the theory, particularly with 99% intervals. In a classroom exercise it is wise to arrange for combined presentation and discussion of the results from the whole class. Preserve the results (sample sizes and numbers of "Yes" answers) since they will be used again later.

You have now done experimentally what the mathematical statistician does theoretically when he studies the distribution of samples drawn at random from a specified population.

For illustration, suppose that an odd digit represents a "Yes" answer, and that the first sample, of size 50, is the first column of table A 1. Counting down the column, you will find 24 odd digits. From table 1.4.1, the 95% confidence interval extends from 36% to 64%, a correct verdict because it includes the population value of 50%. But suppose one of your samples of 250 had started at row 85, column 23. Moving down the successive columns you would count only 101 or 40.4% odd and would assert that the true value is between 34% and 46%. You would be wrong despite the fact that the sample is randomly drawn from the same population as the others. This sample merely happens to be unusually divergent. You should find about five samples in a hundred leading to incorrect statements, but there will be no occasion for surprise if only three, or as many as seven, turn up. With confidence probability 99% you expect, of course, only about one statement in a hundred to be wrong. We hope that your results are sufficiently concordant with theory to give you confidence in it. You will certainly be more aware of the vagaries of sampling, and this is one of the objectives of the experiment. Another lesson to be learned is that only broad confidence intervals can be based on small samples, and that even so the inference can be wrong.

Finally, as is evident in table 1.4.1, you may have observed that the interval narrows rather slowly with increasing sample size. For samples of size 100 that show a percentage of "Yes" answers anywhere between 40% and 60%, the 95% confidence interval is consistently of width 20%.

With a sample ten times as large ($n = 1,000$) the width of the interval decreases to 6%. The width goes down roughly as the square root of the sample size, since $20/6$ is 3.3 and $\sqrt{10}$ is 3.2 (this result was verified in example 1.4.8).

Failure to make correct inferences in a small portion of the samples is not a fault that can be remedied, but a fault inevitably bound up in the sampling procedure. Fallibility is in the very nature of such evidence. The sampler can only take available precautions, then prepare himself for his share of mistakes. In this he is not alone. The journalist, the judge, the banker, the weather forecaster—these along with the rest of us are subject to the laws of chance, and each makes his own quota of wrong guesses. The statistician has this advantage: he can, in favorable circumstances, know his likelihood of error.

1.8—The sampled population. Thus far we have learned that if we want to obtain some information about a population that is too large to be completely studied, one way to do this is to draw a random sample and construct point and interval estimates, as in the Boone County example. This technique of making inferences from sample to population is one of the principal tools in the analysis of data. The data, of course, represent the sample, but the concept of the population requires further discussion. In many investigations in which data are collected, the population is quite specific, apart possibly from some problems of definition: the patients in a hospital on a particular day, the payments received by a firm during the preceding year, and so on. In such cases the investigator often proceeds to select a simple random sample, or one of the more elaborate methods of sampling to be presented in chapter 17, and makes inferences directly from his sample to his population.

With a human population, however, the population actually sampled may be narrower than the original population because some persons drawn into the sample cannot be located, are ill, or refuse to answer the questions asked. Non-responses of this kind in 5% to 15% of the sample are not uncommon. The population to which statistical inferences apply must be regarded as the aggregate of persons who would supply answers if drawn into the sample.

Further, for reasons of feasibility or expense, much research is carried out on populations that are greatly restricted as compared to the population about which, ideally, the investigator would like to gain information. In psychology and education the investigator may concentrate on the students at a particular university, although he hopes to find results that apply to all young men in the country of college age. If the measuring process is troublesome to the person being measured, the research worker may have to depend on paid volunteers. In laboratory research on animals the sample may be drawn from the latest group of animals sent from the supply house. In many of these cases the sampled population, from the viewpoint of statistical inference, is hard to define concretely. It is the kind of population of which the data can be regarded as a random sample.

Confidence interval statements apply to the population that was actually sampled. Claims that such inferences apply to some more extensive population must rest on the judgment of the investigator or on additional extraneous information that he possesses. Careful investigators take pains to describe any relevant characteristics of their data in order that the reader can envisage the nature of the sampled population. The investigator may also comment on ways in which his sampled population appears to differ from some broader population that is of particular interest. As is not surprising, results soundly established in narrow populations are sometimes shown to be erroneous in much broader populations. Fortunately, local studies that claim important results are usually repeated by investigators in other parts of the country or the world, so that a more extensive target population is at least partially sampled in this way.

1.9—The frequency distribution and its graphical representation. One group of students drew 200 samples, each of size 10. The combined results are compactly summarized in a *frequency distribution*, shown in table 1.9.1. There are only eleven possible results for the number of odd digits in a sample, namely the integers 0, 1, 2, . . . 10. Consequently, the frequency distribution has eleven *classes*. The number of samples out of the 200 that fall into a class is the *class frequency*. The sum of the class frequencies is, of course, the total number of samples drawn, 200. The classes and their frequencies give a complete summary of the drawings.

This type of frequency distribution is called *discrete*, because the variable, number of odd digits, can take only a limited number of distinct values. Later we shall meet *continuous* frequency distributions, which are extensively used with measurement data.

One striking feature of the sampling distribution is the concentra-

TABLE 1.9.1
FREQUENCY DISTRIBUTION OF NUMBERS OF ODD DIGITS IN 200 SAMPLES OF $n = 10$

Class (Number of Odd Digits)	Class Frequency	Theoretical Class Frequency
0	1	0.2
1	1	2.0
2	8	8.8
3	25	23.4
4	39	41.0
5	45	49.2
6	36	41.0
7	25	23.4
8	16	8.8
9	4	2.0
10	0	0.2
Total Frequency	200	200.0

tion of frequencies near the middle of the table. The greatest frequency is in the class of five odd digits; that is, half odd and half even. The three middle classes, 4, 5, 6, contain $39 + 45 + 36 = 120$ samples, more than half of the total frequency. This central tendency is the characteristic that gives us confidence in sampling—most samples furnish close estimates of the population fraction of odds. This should counterbalance the perhaps discouraging fact that some of the samples are notably divergent.

Another interesting feature is the symmetry of the distribution, the greatest frequency at the center with a trailing away at each end. This is because the population fraction is 50%; if the percentage were nearer zero or 100, the frequencies would pile up at or near one end.

The regularity that has appeared in the distribution shows that chance events follow a definite law. The turning up of odd digits as you counted them may have seemed wholly erratic: whether an odd or an even would come next was a purely chance event. But the summary of many such events reveals a pattern which may be predicted (aside from sampling variation).

Instead of showing the class frequencies in table 1.9.1, we might have divided each class frequency by 200, the number of samples, obtaining a set of *relative* class frequencies that add to 1. As the number of samples is increased indefinitely, these relative frequencies tend to certain fixed values that can be calculated from the theory of probability. The theoretical distribution computed in this way is known as the *binomial distribution*. It is one of the commonest distributions in statistical work. In general terms, the formula for the binomial distribution is as follows. Suppose that we are drawing samples of size n and that the attribute in question is held by a proportion p of the members of the population. The relative frequency of samples containing r members having the attribute, or in other words the probability that a sample will contain r members having the attribute, is

$$\frac{n(n-1)(n-2) \dots (n-r+1)}{r(r-1)(r-2) \dots (2)(1)} p^r (1-p)^{n-r}$$

In the numerator the expression $n(n-1)(n-2) \dots (n-r+1)$ means "multiply together all the integers from n down to $(n-r+1)$, inclusive." Similarly, the first term in the denominator is a shorthand way of writing the instruction "multiply together all integers from r down to 1." We shall study the binomial distribution and its mathematical derivation in chapter 8.

What does this distribution look like for our sampling in table 1.9.1? We have $n = 10$ and $p = 1/2$. The relative frequency or probability of a sample having four odd digits is, putting $r = 4$ so that $(n-r+1) = 7$,

$$\frac{(10)(9)(8)(7)(1/2)^4 (1/2)^6}{(4)(3)(2)(1)} = (210) \left(\frac{1}{2}\right)^{10} = \frac{210}{1024}$$

As already mentioned, these relative frequencies add to 1. (This is not obvious by looking at the formula, but comes from a well-known result in algebra.) Hence, in our 200 samples of size 10, the number that should theoretically have four odd digits is

$$\frac{(200)(210)}{1024} = 41.0$$

These theoretical class frequencies are given in the last column of table 1.9.1. The agreement between the actual and theoretical frequencies is pleasing.

The graph in figure 1.9.1 brings out the features of the binomial distribution. On the horizontal axis are marked off the different classes—the numbers of odd digits. The solid ordinate beside each class number is the observed class frequency while the dotted ordinate represents the theoretical frequency. This is the type of graph appropriate for discrete distributions.

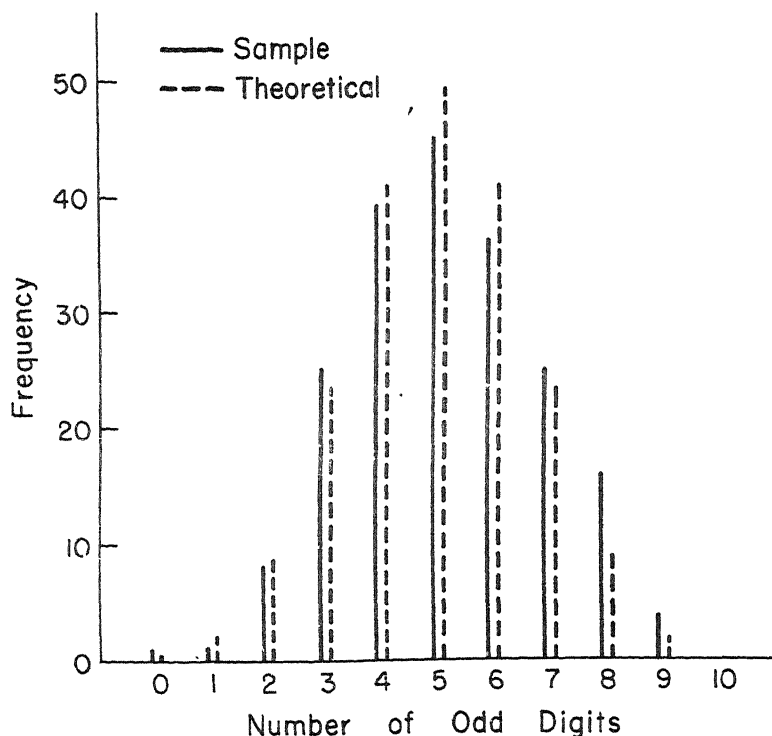


FIG. 1.9.1—Frequency distribution of number of odd digits in each of 200 samples of size 10. The dotted lines represent the theoretical binomial distribution from which the samples were drawn.

EXAMPLE 1.9.1—For the 200 samples of size 10 in table 1.9.1, in how many cases is (i) the 95% confidence interval statement wrong? (ii) the 99% confidence interval statement wrong? Ans. (i) 6 times, or 3.0%; (ii) 1 time, or 0.5%.

EXAMPLE 1.9.2—Use the table of random digits to select a random sample of 20 pages of this book, regarding the population as consisting of pages 3–539. Note the number of pages in your sample that do not contain the beginning of a new section, and calculate the 95% interval for the proportion of pages in the book on which no new section begins. Don't count "References" as a section. The population proportion is $317/537 = 0.59$.

EXAMPLE 1.9.3—When the doors of a clinic are opened, twelve patients enter simultaneously. Each patient wishes to be handled first. Can you use the random digit table to arrange the patients in a random order?

EXAMPLE 1.9.4—A sampler of public opinion estimates from a sample the number of eligible voters in a state favoring a certain candidate for governor. Assuming that his estimate was close to the population value at the time the survey was made, suggest two reasons why the ballot on election day might be quite different.

EXAMPLE 1.9.5—A random sample of families from a population has been selected. An interviewer calls on each family at its home between the hours of 9 A.M. and 5 P.M. If no one is at home, the interviewer makes no attempt to contact the family at a later time. For each of the following attributes, give your opinion whether the sample results are likely to overestimate, underestimate, or be at about the correct level: (i) proportion of families in which the husband is retired, (ii) proportion of families with at least one child under 4 years, (iii) proportion of families in which husband and wife both work. Give your reasons.

EXAMPLE 1.9.6—From the formula for the binomial distribution, calculate the probability of 0, 1, 2 "Yes" answers in a sample of size 2, where p is the proportion of "Yes" answers in the population. Show that the three probability values add to 1 for any value of p .

EXAMPLE 1.9.7—At birth the probability that a child is a boy is very close to one-half. Show that according to the binomial distribution, half the families of size 2 should consist of one boy and one girl. Why is the proportion of boy-girl families likely to be slightly less than one-half in practice?

EXAMPLE 1.9.8—Five dice were tossed 100 times. At each toss the number of two's (deuces) out of five were noted, with these results:

Number Deuces Per Toss	Frequency of Occurrence	Theoretical Frequency
5	2	0.013
4	3	0.322
3	3	3.214
2	18	16.075
1	42	40.188
0	32	40.188
Total	100	100.000

(i) From the binomial distribution, verify the result 16.075 for the theoretical frequency of 2 deuces. (ii) Draw a graph showing the observed and theoretical distributions. (iii) Do you think the dice were balanced and fairly tossed? Ans. The binomial probability of 2 deuces is $1250/7776 = 0.16075$. This is multiplied by 100 to give the theoretical frequency. A later test (example 9.5.1) casts doubt on the throws.

1.10—Hypotheses about populations. The investigator often has in mind a definite hypothesis about the population ratio, the purpose of the sampling being to get evidence concerning his hypothesis. Thus a geneticist studying heredity in the tomato had reason to believe that in the plants produced from a certain cross, fruits with red flesh and yellow flesh would be in the ratio 3:1. In a sample of 400 he found 310 red tomatoes instead of the hypothetical 300. With your experience of sampling variation, would you accept this as verification or refutation of the hypothesis? Again, a physician has the hypothesis that a certain disease requiring hospitalization is equally common among men and women. In a sample of 900 hospital cases he finds 480 men and 420 women. Do these results support or contradict his hypothesis? (Incidentally, this is an example in which the sampled population may differ from the target population. Although good medical practice may prescribe hospitalization, there are often cases that for one reason or another do not come to a hospital and therefore could not be included in his sample.)

To answer such questions two results are needed, a measure of the deviation of the sample from the hypothetical population ratio, and a means of judging whether this measure is an amount that would commonly occur in sampling, or, on the contrary, is so great as to throw doubt upon the hypothesis. Both results were furnished by Karl Pearson in 1899 (3). He devised an index of dispersion or test criterion denoted by χ^2 (chi-square) and obtained the formula for its theoretical frequency distribution when the hypothesis in question is true. Like the binomial distribution, the chi-square distribution is another of the basic theoretical distributions much used in statistical work. Let us first examine the index of dispersion.

1.11—Chi-square, an index of dispersion. Naturally, the deviations of the observed numbers from those specified by the hypothesis form the basis of the index. In the medical example, with 900 cases, the numbers of male and female cases expected on the hypothesis are each 450. The deviations, then, are

$$480 - 450 = +30,$$

and

$$420 - 450 = -30,$$

the sum of the two being zero. The value of chi-square is given by

$$\chi^2 = \frac{(+30)^2}{450} + \frac{(-30)^2}{450} = 2 + 2 = 4$$

Each deviation is squared, each square is divided by the hypothetical or expected number, and the results are added. The expected numbers appear in the denominators in order to introduce sample size into the quantity—it is the *relative* size that is important.

The squaring of the deviations in the numerator may puzzle you.

It is a common practice in statistics. We shall simply say at present that indexes constructed in this way have been found to have great flexibility, being applicable to many different types of statistical data. Note that the squaring makes the sign of the deviation unimportant, since the square of a negative number is the same as that of the corresponding positive number. It is clear that chi-square would be zero if the sample frequencies were the same as the hypothetical, and that it will increase with increasing deviation from the hypothetical. But it is not at all clear whether a chi-square value of 4 is to be considered large, medium, or small.

To furnish a basis for judgment on this point is our next aim. Pearson founded his judgment from a study of the theoretical distribution of chi-square, but we shall investigate the same problem by setting up a sampling experiment. Before doing this, a useful formula will be given, together with a few examples to help fix it in mind.

1.12—The formula for chi-square. It is convenient to represent by f_1 and f_2 the sample counts of individuals who do and do not possess the attribute being investigated, the corresponding hypothetical or expected frequencies being F_1 and F_2 . The two deviations, then, are $f_1 - F_1$ and $f_2 - F_2$, so that chi-square is given by the formula,

$$\chi^2 = (f_1 - F_1)^2/F_1 + (f_2 - F_2)^2/F_2$$

The formula may be condensed to the more easily remembered as well as more general one,

$$\chi^2 = \Sigma(f - F)^2/F,$$

where Σ denotes summation. In words, "Chi-square is the sum of such ratios as

(deviation squared)/(expected number)"

Let us apply the formula to the counts of red and yellow tomatoes in section 1.10. There, $f_1 = 310$, $f_2 = 400 - 310 = 90$, $F_1 = 3/4$ of $400 = 300$, and $F_2 = 1/4$ of $400 = 100$. Whence,

$$\chi^2 = \frac{(310 - 300)^2}{300} + \frac{(90 - 100)^2}{100} = 1.33$$

Note. When computing chi-square it is essential to use the actual size of sample and the actual numbers in the two attribute classes. If we know only the percentages or proportions in the two classes, *chi-square cannot be calculated*. Suppose we are told that 80% of the tomato plants in a sample are red, and asked to compute chi-square. If we guess that the sample contained 100 plants then

$$\chi^2 = \frac{(80 - 75)^2}{75} + \frac{(20 - 25)^2}{25} = \frac{25}{75} + \frac{25}{25} = 1.33$$

22 Chapter 1: Sampling of Attributes

But if the sample actually contained only 10 plants, then

$$\chi^2 = \frac{(8 - 7.5)^2}{7.5} + \frac{(2 - 2.5)^2}{2.5} = \frac{0.25}{7.5} + \frac{0.25}{2.5} = 0.133$$

If the sample had 1,000 plants, a similar calculation finds $\chi^2 = 13.33$. For a given percentage red, the value of chi-square can be anything from almost zero to a very large number.

EXAMPLE 1.12.1—A student tossed a coin 800 times, getting 440 heads. What is the value of chi-square in relation to the hypothesis that heads and tails are equally likely? Ans. 8.

EXAMPLE 1.12.2—If the count in the preceding example had been 220 heads out of 400 tosses, would chi-square also be half its original value?

EXAMPLE 1.12.3—A manufacturer of a small mass-produced article claims that 96% of the articles function properly. In an independent test of 1,000 articles, 950 were found to function properly. Compute chi-square. Ans. 2.60.

EXAMPLE 1.12.4—In the text example about tomatoes the deviation from expectation was 10. If the same deviation had occurred in a sample of twice the size (that is, of 800), what would have been the value of chi-square? Ans. 0.67, half the original value.

1.13—An experiment in sampling chi-square; the sampling distribution.

You have now had some practice in the calculation of chi-square. Its main function is to enable us to judge whether the sample ratio itself departs much or little from the hypothetical population value. For that purpose we must answer the question already proposed: What values of chi-square are to be considered as indicating unusual deviation, and what as ordinary sampling variation? Our experimental method of answering the question will be to calculate chi-square for each of many samples drawn from the table of random numbers, then to observe what values of chi-square spring from the more unusual samples. If a large number of samples of various sizes have been drawn and if the value of chi-square is computed from each, the distribution of chi-square may be mapped.

The results to be presented here come from 230 samples of sizes varying from 10 to 250, drawn from the random digits table A 1. We suggest that the reader use the samples that he drew in section 1.7 when verifying the confidence interval statements. There is a quick method of calculating chi-square for all samples of a given size n . Since odd and even digits are equally likely in the population, the expected numbers of odd and even digits are $F_1 = F_2 = n/2$. The reciprocals of these numbers are therefore both equal to $2/n$. Remembering that the two deviations are the same in absolute value and differ only in sign, we may write

$$\begin{aligned}\chi^2 &= (f_1 - F_1)^2(1/F_1 + 1/F_2) \\ &= d^2(2/n + 2/n) = 4d^2/n\end{aligned}$$

where d is the absolute value of the deviation. For all samples of a fixed size n , the multiplier $4/n$ is constant. Once it has been calculated it can be used again and again.

To illustrate, suppose that $n = 100$. The multiplier $4/n$ is 0.04. If 56 odd digits are found in a sample, $d = 6$ and

$$\chi^2 = (0.04)(6^2) = 1.44$$

Proceed to calculate chi-square for each of your samples. To summarize the results, a frequency distribution is again convenient. There is one difference, however, from the discrete frequency distribution used in section 1.9 when studying the binomial distribution. With the binomial for $n = 10$, there were only eleven possible values for the numbers of odd digits, so that the eleven classes in the frequency distribution selected themselves naturally. On the other hand, with chi-square values calculated from samples of different sizes, there is a large number of possible values. Some grouping of the values into classes is necessary. A distribution of this type is sometimes described as *continuous*, since conceptually any positive number is a possible value of chi-square.

When forming frequency distributions from continuous data, decide first on the classes to be used. For most purposes, somewhere between 8 and 20 classes is satisfactory. Obtain an idea of the range of the data by looking through them quickly to spot low and high values. Most of your chi-squares will be found to lie between 0 and 5. Equal-sized class intervals of 0.00–0.49, 0.50–0.99, . . . will therefore cover most of the range in 10 classes, although a few values of chi-square greater than 5 may occur. Our values of χ^2 were recorded to 2 decimal places.

Be sure to make the classes non-overlapping, and indicate clearly what the class intervals are. Class intervals described as “0.00–0.50,” “0.50–1.00,” “1.00–1.50” are not satisfactory, since the reader does not know in what classes the values 0.50 and 1.00 have been placed. If the chi-square values were originally computed to *three* decimal places, reported class intervals of “0.00–0.49,” “0.50–0.99,” and so on, would be

TABLE 1.13.1
SAMPLING DISTRIBUTION OF 230 VALUES OF CHI-SQUARE CALCULATED FROM SAMPLES
DRAWN FROM TABLE A 1
Sample sizes— 10, 15, 20, 30, 50, 100, and 250

Class Interval	Frequency	Class Interval	Frequency
0.00–0.49	116	6.00– 6.49	0
0.50–0.99	39	6.50– 6.99	1
1.00–1.49	18	7.00– 7.49	0
1.50–1.99	22	7.50– 7.99	0
2.00–2.49	12	8.00– 8.49	0
2.50–2.99	5	8.50– 8.99	1
3.00–3.49	5	9.00– 9.49	0
3.50–3.99	6	9.50– 9.99	0
4.00–4.49	1	10.00–10.49	1
4.50–4.99	2	10.50–10.99	0
5.00–5.49	0	11.00–11.49	1
5.50–5.99	0	Total	230

ambiguous, since it is not clear where a chi-square value of 0.493 is placed. Intervals of 0.000–0.494, 0.495–0.999, and so on, could be used.

Having determined the class intervals, go through the data systematically, assigning each value of chi-square to its proper class, then counting the number of values (frequency) in each class. Table 1.13.1 shows the results for our 230 samples.

In computing chi-square, we chose to regard the population as consisting of the 10,000 random digits in table A 1, rather than as an infinite population of random digits. Since 5,060 of the digits in table A 1 are odd, we took the probability of an odd digit as 0.506 instead of 0.50. The reader is recommended to use 0.50, as already indicated. The change makes only minor differences in the distribution of the sample values of chi-square.

Observe the concentration of sample chi-squares in the smallest class, practically half of them being less than 0.5. Small deviations (with small chi-squares) are predominant, this being the foundation of our faith in sampling. But taking a less optimistic view, one must not overlook the samples with large deviations and chi-squares. The possibility of getting one of these makes for caution in drawing conclusions. In this sampling exercise we know the population ratio and are not led astray by discrepant samples. In actual investigations, where the hypothesis set up is not known to be the right one, a large value of chi-square constitutes a dilemma. Shall we say that it denotes only an unusual sample from the hy-

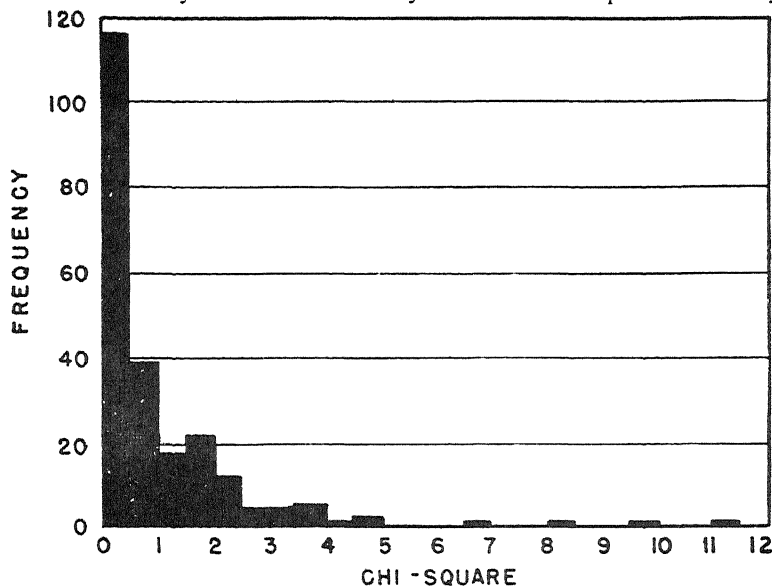


FIG 1.13.1—Histogram representing frequency distribution of the 230 sample values of chi-square in table 1.13.1

pothetical population, or shall we conclude that the hypothesis misrepresents the true population ratio? Statistical theory contains no certain answer. Instead, it furnishes an *evaluation of the probability of possible sample deviations from the hypothetical population*. If chi-square is large, the investigator is warned that the sample is an improbable one under his hypothesis. This is evidence to be added to that which he already possesses, all of it being the basis for his decisions. A more exact determination of probability will be explained in section 1.15.

The graphical representation of the distribution of our chi-squares appears in figure 1.13.1. In this kind of graph, called a *histogram*, the frequencies are represented by the areas of the rectangular blocks in the figure. The graph brings out both the concentration of small chi-square at the left and the comparatively large sizes of a few at the right. It is now evident that for the medical example in section 1.11, $\chi^2 = 4$ is larger than a great majority of the chi-squares in this distribution. If this disease were in fact equally likely to result in male or female hospitalized cases, this would be an unusually large value of chi-square.

1.14—Comparison with the theoretical distribution. Two features of our chi-square distribution have yet to be examined: (i) How does it compare with the theoretical distribution? and (ii) How can we evaluate more exactly the probabilities of various chi-square sizes? For these purposes a rearrangement of the class intervals is advisable. Since our primary interest is in the relative frequency of high values of chi-square, we used the set of class intervals defined by column 4 of table 1.14.1. The first three intervals each contain 25% of the theoretical distribution. As chi-square increases, the next four intervals contain respectively 15%, 5%, 4%, and

TABLE 1.14.1
COMPARISON OF THE SAMPLE AND THEORETICAL DISTRIBUTIONS OF CHI-SQUARE

Class Interval of Chi-square	Sample Frequency Distribution		Theoretical Frequency Distribution		
	Actual	Percentage	Percentage	Cumulative χ^2	Per Cent Greater Than
1	2	3	4	5	6
0-0.1015	57	24.8	25	0	100
0.1015-0.455	59	25.6	25	0.1015	75
0.455-1.323	62	27.0	25	0.455	50
1.323-2.706	32	13.9	15	1.323	25
2.706-3.841	14	6.1	5	2.706	10
3.841-6.635	3	1.3	4	3.841	5
6.635-	3	1.3	1	6.635	1
Total	230	100.0	100		

1%. Since the theoretical distribution is known exactly and has been widely tabulated, the corresponding class intervals for chi-square, shown in column 1, are easily obtained. Note that the intervals are quite unequal.

Column 2 of table 1.14.1 shows the actual frequencies obtained from the 230 samples. In column 3, these have been converted to percentage frequencies, by multiplying by $100/230$, for comparison with the theoretical percentage frequencies in column 4. The agreement between columns 3 and 4 is good. If your chi-square values have been computed mostly from small samples of sizes 10, 15, and 20, your agreement may be poorer. With small samples there is only a limited number of distinct values of chi-square, so that your sample distribution goes by discontinuous jumps.

Columns 5 and 6 contain a cumulative frequency distribution of the percentages in column 4. Beginning at the foot of column 6, each entry is the sum of all the preceding ones in column 4, hence the name. The column is read in this way: the third to the last entry means that 10% of all samples in the theoretical distribution have chi-squares greater than the 2.706. Again, 50% of them exceed 0.455; this may be looked upon as an average value, exceeded as often as not in the sampling. Finally, chi-squares greater than 6.635 are rare, occurring only once per 100 samples. So in this sampling distribution of chi-square we find a measure in terms of probability, the measure we have been seeking to enable us to say exactly which chi-squares are to be considered small and which large. We are now to learn how this measure can be utilized.

1.15—The test of a null hypothesis or test of significance. As indicated in section 1.10, the investigator's objective can often be translated into a hypothesis about his experimental material. The geneticist, you remember, knowing that the Mendelian theory of inheritance produced a 3:1 ratio, set up the hypothesis that the tomato population had this ratio of red to yellow fruits. This is called a *null hypothesis*, meaning that there is no difference between the hypothetical ratio and that in the population of tomato fruits. If this null hypothesis is true, then random samples of n will have ratios distributed binomially, and chi-squares calculated from the samples will be distributed as in table 1.14.1. To *test the hypothesis*, a sample is taken and its chi-square calculated; in the illustration the value was 1.33. Reference to the table shows that, if the null hypothesis is true, 1.33 is not an uncommon chi-square, the probability of a greater one being about 0.25. As the result of this test, the geneticist would not likely reject the null hypothesis. He knows, of course, that he may be in error, that the population ratio among the tomato fruits may not be 3:1. But the discrepancy, if any, is so small that the sample has given no convincing evidence of it.

Contrasting with the genetic experiment, the medical example turned up $\chi^2 = 4$. If the null hypothesis (this disease equally likely in men and women) is true, a larger chi-square has a probability of only about 0.05. This suggests that the null hypothesis is false, so the sampler would likely

reject it. As before, he may be in error because this might be one of those 5 samples per 100 that have chi-squares greater than 3.841 even when the sampling is from an equally divided population. In rejecting the null hypothesis, the sampler faces the possibility that he is wrong. Such is the risk always run by those who test hypotheses and rest decisions on the tests.

The illustrations show that in testing hypotheses one is liable to *two kinds of error*. If his sample leads him to reject the null hypothesis when it is true, he is said to have committed an *error of the first kind*, or a Type I error. If, on the contrary, he is led to accept the hypothesis when it is false, his *error is of the second kind*, a Type II error. The Neyman-Pearson theory of testing hypotheses emphasizes the relations between these types. For recent accounts of this theory see references (6, 7, 8).

As a matter of practical convenience, probability levels of 5% (0.05) and 1% (0.01) are commonly used in deciding whether to reject the null hypothesis. As seen from table 1.14.1, these correspond to χ^2 greater than 3.841 and χ^2 greater than 6.635, respectively. In the medical example we say that the difference in the number of male and female patients is *significant at the 5% level*, because it signifies rejection of the null hypothesis of equal numbers.

This use of 5% and 1% levels is simply a working convention. There is merit in the practice, followed by some investigators, of reporting in parentheses the probability that chi-square exceeds the value found in their data. For instance, in the counts of red and yellow tomatoes, we found $\chi^2 = 1.33$, a value exceeded with probability about 0.25. The report might read: "The χ^2 test was consistent with the hypothesis of a 3 to 1 ratio of red to yellow tomatoes ($P = 0.25$)."

The values of χ^2 corresponding to a series of probability levels are shown below. This table should be used in working the exercises that follow.

Probability of a Greater Value									
P	0.90	0.75	0.50	0.25	0.10	0.05	0.025	0.010	0.005
χ^2	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88

EXAMPLE 1.15.1—Two workers A and B perform a task in which carelessness leads to minor accidents. In the first 20 accidents, 13 happened to A and 7 to B. Is this evidence against the hypothesis that the two men are equally liable to accidents? Compute χ^2 and find the significance probability. Ans. $\chi^2 = 1.8$. P between 0.10 and 0.25.

EXAMPLE 1.15.2—A baseball player has a lifetime batting average of 0.280. (This means that the probability that he gets a hit when at bat is 0.280.) Starting a new season, he gets 15 hits in his first 30 times at bat. Is this evidence that he is having what is called a hot streak? Compute χ^2 for the null hypothesis that his probability of hitting is still 0.280. Ans. $\chi^2 = 7.20$. $P < 0.01$. Null hypothesis is rejected.

EXAMPLE 1.15.3—In some experiments on heredity in the tomato, MacArthur (5) counted 3,629 fruits with red flesh and 1,176 with yellow. This was in the F_2 generation

where the theoretical ratio was 3 : 1. Compute $\chi^2 = 0.71$ and find the significance probability. MacArthur concluded that "the discrepancies between the observed and expected ratios are not significant."

EXAMPLE 1.15.4—In a South Dakota farm labor survey of 1943, 480 of the 1,000 reporting farmers were classed as owners (or part owners), the remaining 520 being renters. It is known that of nearly 7,000 farms in the region, 47% are owners. Assuming this to be population percentage, calculate chi-square and P for the sample of 1,000. Ans. $\chi^2 = 0.41$, $P = 0.50$. Does this increase your confidence in the randomness of the sampling? Such collateral evidence is often cited. The assumption is that if the sample is shown to be representative for one attribute it is more likely to be representative also of the attribute under investigation, provided the two are related.

EXAMPLE 1.15.5—James Snedecor (4) tried the effect of injecting poultry eggs with female sex hormones. In one series 2 normal males were hatched together with 19 chicks which were classified as either normal females or as individuals with pronounced female characteristics. What is the probability of the ratio 2 : 19, or one more extreme, in sampling from a population with equal numbers of the sexes in which the hormone has no effect? Ans. $\chi^2 = 13.76$, P is much less than 0.01.

EXAMPLE 1.15.6—In table 1.14.1, there are $62 + 32 + 14 + 3 + 3 = 114$ samples having chi-squares greater than 0.455, whereas 50% or 230 were expected. What is the probability of drawing a more discrepant sample if the sampling is truly random? Ans. $\chi^2 = 0.0174$, $P = 0.90$. Make the same test for your own samples.

EXAMPLE 1.15.7—This example illustrates the discontinuity in the distribution of chi-square when computed from small samples. From 100 samples of size 10 drawn from the random digits table A 1, the following frequency distribution of the numbers of odd digits in a sample was obtained.

Number of odd digits	1 or 9	2 or 8	3 or 7	4 or 6	5
Frequency	2	8	19	46	25

Compute the sample frequency distribution of χ^2 as in table 1.14.1 and compare it with the theoretical distribution. Observe that no sample χ^2 occurs in the class interval 0.455–1.323, although 25% of the theoretical distribution lies in this range.

1.16—Tests of significance in practice. A test of significance is sometimes thought to be an automatic rule for making a decision either to "accept" or "reject" a null hypothesis. This attitude should be avoided. An investigator rarely rests his decisions wholly on a test of significance. To the evidence of the test he adds knowledge accumulated from his own past work and from the work of others. The size of the sample from which the test of significance is calculated is also important. With a small sample, the test is likely to produce a significant result only if the null hypothesis is very badly wrong. An investigator's report on a small sample test might read as follows: "Although the deviation from the null hypothesis was not significant, the sample is so small that this result gives only a weak confirmation of the null hypothesis." With a large sample, on the other hand, small departures from the null hypothesis can be detected as statistically significant. After comparing two proportions in a large sample, an investigator may write: "Although statistically significant, the difference between the two proportions was too small to be of practical importance, and was ignored in the subsequent analysis."

In this connection, it is helpful, when testing a binomial proportion at the 5% level, to look at the 95% confidence limits for the population p . Suppose that in the medical example the number of patients was only $n = 10$, of whom 4 were female, so that the sample proportion of female patients was 0.4. If you test the null hypothesis $p = 0.5$ by χ^2 , you will find $\chi^2 = 0.4$, a small value entirely consistent with the null hypothesis. Looking now at the 95% confidence limits for p , we find from table 1.4.1 (p. 000) that these are 15% and 74%. Any value of the population p lying between 15% and 74% is also consistent with the sample result. Clearly, the fact that we found a non-significant result when testing the null hypothesis $p = 1/2$ gives no assurance from these data that the true p is $1/2$ or near to $1/2$.

1.17—Summary of technical terms. In this chapter you have been introduced to some of the main ideas in statistics, as well as to a number of the standard technical terms. As a partial review and an aid to memory, these terms are described again in this section. Since these descriptions are not dictionary definitions, some would require qualification from a more advanced viewpoint, but they are substantially correct.

Statistics deals with techniques for collecting, analyzing, and drawing conclusions from data.

A *sample* is a small collection from some larger aggregate (the *population*) about which we wish information.

Statistical inference is concerned with attempts to make quantitative statements about properties of a population from a knowledge of the results given by a sample.

Attribute data are data that consist of a classification of the members of the sample into a limited number of classes on the basis of some property of the members (for instance, hair color). In this chapter, only samples with two classes have been studied.

Measurement data are data recorded on some numerical scale. They are called *discrete* when only a restricted number of values occurs (for instance, 0, 1, 2, . . . 11 children). Strictly, all measurement data are discrete, since the results of any measuring process are recorded to a limited number of figures. But measurement data are called *continuous* if, conceptually successive values would differ only by tiny amounts.

A *point estimate* is a single number stated as an estimate of some quantitative property of the population (for instance, 2.7% defective articles, 58,300 children under five years). The quantity being estimated is often called a *population parameter*.

An *interval estimate* is a statement that a population parameter has a value lying between two specified limits (the population contains between 56,900 and 60,200 children under five years).

A *confidence interval* is one type of interval estimate. It has the feature that in repeated sampling a known proportion (for instance, 95%) of the intervals computed by this method will include the population parameter.

Random sampling, in its simplest form, is a method of drawing a sample such that any member of the population has an equal chance of appearing in the sample, independently of the other members that happen to fall in the sample.

Tables of random digits are tables in which digits 0, 1, 2, . . . 9 have been drawn by some process that gives each digit an equal chance of being selected at any draw.

The *sampled population* is the population of which our data are a random sample. It is an aggregate such that the process by which we obtained our sample gives every member of the aggregate a known chance of appearing in the sample, and is the population to which statistical inferences from the sample apply. In practice, the sampled population is sometimes hypothetical rather than real, because the only available data may not have been drawn at random from a known population. In meteorological research, for instance, the best data might be weather records for the past 40 years, which are not a randomly selected sample of years.

The *target population* is the aggregate about which the investigator is trying to make inferences from his sample. Although this term is not in common use, it is sometimes helpful in focussing attention on differences between the population actually sampled and the population that we are attempting to study.

In a *frequency distribution*, the values in the sample are grouped into a limited number of classes. A table is made showing the class boundaries and the frequencies (number of members of the sample) in each class. The purpose is to obtain a compact summary of the data.

The *binomial distribution* gives the probabilities that 0, 1, 2, . . . n members of a sample of size n will possess some attribute, when the sample is a random sample from a population in which a proportion p of the members possess this attribute.

A *null hypothesis* is a specific hypothesis about a population that is being tested by means of the sample results. In this chapter the only hypothesis considered was that the proportion of the population having some attribute has a stated numerical value.

A *test of significance* is, in general terms, a calculation by which the sample results are used to throw light on the truth or falsity of a null hypothesis. A quantity called a *test criterion* is computed: it measures the extent to which the sample departs from the null hypothesis in some relevant aspect. If the value of the test criterion falls beyond certain limits into a *region of rejection*, the departure is said to be *statistically significant* or, more concisely, *significant*. Tests of significance have the property that if the null hypothesis is true, the probability of obtaining a significant result has a known value, most commonly 0.05 or 0.01. This probability is the *significance level* of the test.

$\text{Chi-square} = \Sigma (\text{Observed} - \text{Expected})^2 / (\text{Expected})$ is a test criterion for the null hypothesis that the proportion with some attribute in the

population has a specified value. Large values of chi-square are significant. The chi-square criterion serves many purposes and will appear later for testing other null hypotheses.

Errors of the first and second kinds. In the Neyman-Pearson theory of tests of hypotheses, an error of the first kind is the rejection of the null hypothesis when it is true, and an error of the second kind is the acceptance of a null hypothesis that is false. In practice, in deciding whether to reject a null hypothesis or to regard it as provisionally true, all available evidence should be reviewed as well as the specific result of the test of significance.

REFERENCES

1. The confidence intervals for sample sizes up to $n = 30$ were taken from the paper by E. L. Crow, *Biometrika*, 43, 423-435 (1956). Intervals for n greater than 30 were obtained from the normal approximation as discussed in section 8.7.
2. RAND CORPORATION. *A Million Random Digits With 100,000 Normal Deviates*, Free Press, Glencoe, Ill. (1955).
3. K. PEARSON. *Phil. Mag.*, Ser. 5, 50:157 (1899).
4. J. G. SNEDECOR. *J. Exp. Zool.*, 110:205 (1949).
5. J. W. MACARTHUR. *Trans. Roy. Canadian Inst.*, 18:1 (1931).
6. P. G. HOEL. *Introduction to Mathematical Statistics*, 2nd ed., Chap. 10. Wiley, New York (1954).
7. E. S. KEEPING. *Introduction to Statistical Inference*, Chap. 6. Van Nostrand, Princeton, N.J. (1962).
8. H. FREEMAN. *Introduction to Statistical Inference*, Chap. 28. Addison-Wesley, Reading, Mass. (1963).

Sampling from a normally distributed population

2.1—Normally distributed population. In the first chapter, sampling was mostly from a population with only two kinds of individuals; odd or even, alive or dead, infested or free. Random samples of n from such a population made up a *binomial distribution*. The variable, an enumeration of successes, was discrete. Now we turn to another kind of population whose individuals are measured for some characteristic such as height or yield or income. The variable flows without a break from one individual to the next—a continuous variable with no limit to the number of individuals with different measurements. Such variables are distributed in many ways, but we shall be occupied first with the *normal distribution*.

Next to the binomial, the normal distribution was the earliest to be developed. De Moivre published its equation in 1733, twenty years after Bernoulli had given a comprehensive account of the binomial. That the two are not unrelated is clear from figure 2.1.1. On the top is the graph of a symmetrical binomial distribution similar to that in figure 1.9.1. In this new figure the sample size is 48 and the population sampled has equal numbers of the two kinds of individuals. Although discrete, the binomial is here graphed as a histogram. That is, the ordinate at 25 successes is represented by a horizontal bar going from 24.5 to 25.5. This facilitates comparison with the continuous normal curve. An indefinitely great number of samples were drawn so that the frequencies are expressed as percentages of the total. Successes less than 13 and more than 35 do occur, but their frequencies are so small that they cannot be shown on the graph.

Imagine now that the size of the sample is increased without limit, the width of the intervals on the horizontal axis being decreased correspondingly. The steps of the histogram would soon become so small as to look like the continuous curve at the right. Indeed, De Moivre discovered the normal distribution when seeking an approximation to the binomial. The discrete variable has become *continuous* and the frequencies have merged into each other without a break.

This normal distribution is completely determined by two constants or *parameters*. First, there is the *mean*, μ , which locates the center of the distribution. Second, the *standard deviation*, σ , measures the spread or

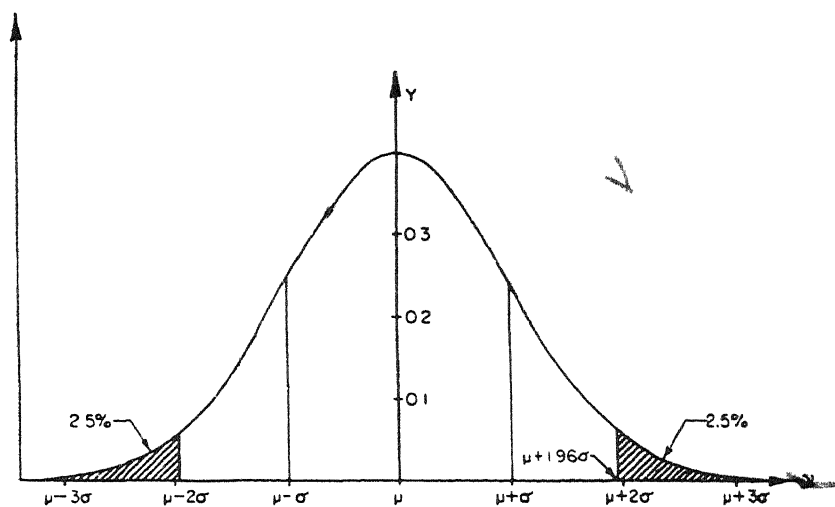
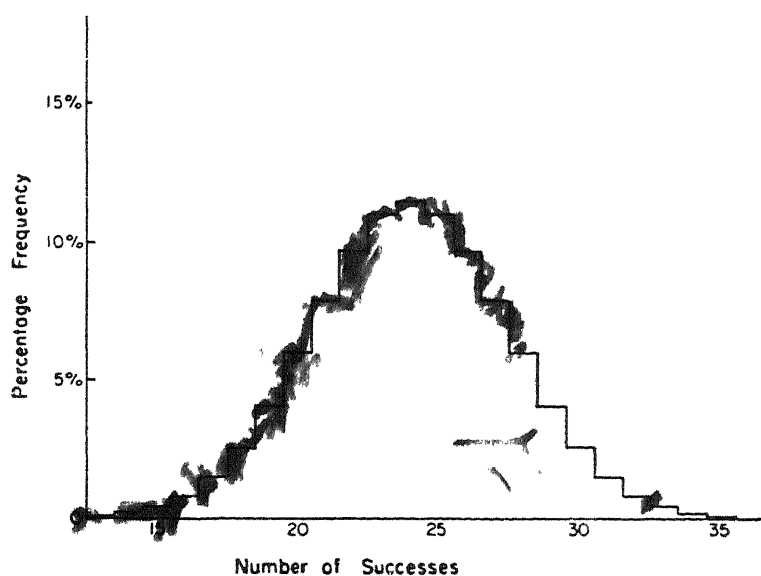


FIG. 2.1.1—Upper binomial distribution of successes in samples of 48 from 1:1 population. Lower normal distribution with mean μ and standard deviation σ ; the shaded areas comprise 5% of the total.

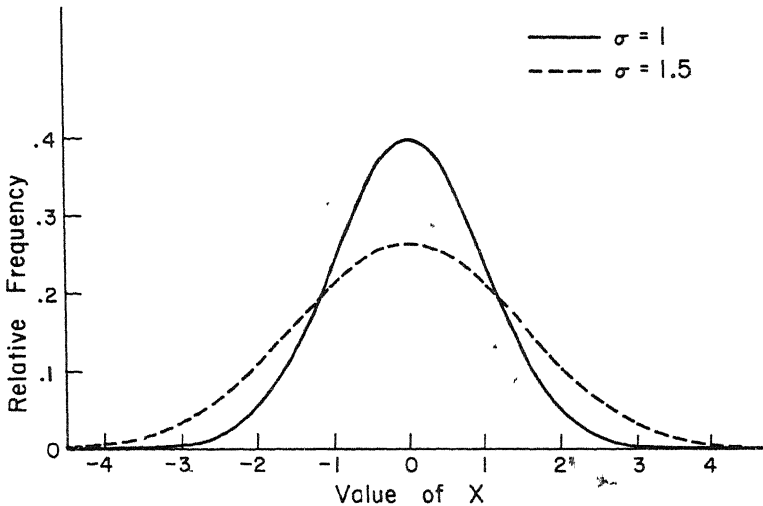


FIG 2.1.2—Solid curve: the normal distribution with $\mu = 0$ and $\sigma = 1$. Dotted curve: the normal distribution with $\mu = 0$ and $\sigma = 1.5$.

variation of the individual measurements; in fact, σ is the *scale* (unit of measurement) of the variable which is normally distributed

From the figure you see that within one sigma on either side of μ the frequency is decreasing ever more rapidly but beyond that point it decreases at a continuously lesser rate. By the time the variable, X , has reached $\pm 3\sigma$ the percentage frequencies are negligibly small. Theoretically, the frequency of occurrence never vanishes entirely, but it approaches zero as X increases indefinitely. The concentration of the measurements close to μ is emphasized by the fact that over $2/3$ of the observations lie in the interval $\mu \pm \sigma$ while some 95% of them are in the interval $\mu \pm 2\sigma$. Beyond $\pm 3\sigma$ lies only 0.26% of the total frequency.

The formula for the ordinate or height of the normal curve is

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2},$$

where the quantity $e = 2.3026$ is the base for natural logarithms and π is of course 3.1416. To illustrate the role of the standard deviation σ in determining the shape of the curve, figure 2.1.2 shows two curves. The solid curve has $\mu = 0$, $\sigma = 1$, while the dotted curve has $\mu = 0$, $\sigma = 1.5$. The curve with the larger σ is lower at the mean and more spread out. Values of X that are far from the mean are much more frequent with $\sigma = 1.5$ than with $\sigma = 1$. In other words, the population is more variable with $\sigma = 1.5$. A curve with $\sigma = 1/2$ would have a height of nearly 0.8 at the mean and would have scarcely any frequency beyond $X = 1.5$.

To indicate the effect of a change in the mean μ , the curve with $\mu = 2$, $\sigma = 1$ is obtained by lifting the solid curve bodily and centering it at $X = 2$ without changing its shape in any other way. This explains why μ is called the parameter of location.

2.2—Reasons for the use of the normal distribution. You may be wondering why such a model is presented since it obviously cannot describe any real population. It is astonishing that this normal distribution has dominated statistical practice as well as theory. Briefly, the main reasons are as follows:

1. Convenience certainly plays a part. The normal distribution has been extensively and accurately tabulated, including many auxiliary results that flow from it. Consequently if it seems to apply fairly well to a problem, the investigator has many time-saving tables ready at hand.

2. The distributions of some variables are approximately normal, such as heights of men, lengths of ears of corn, and, more generally, many linear dimensions, for instance those of numerous manufactured articles.

3. With measurements whose distributions are not normal, a simple transformation of the scale of measurement may induce approximate normality. The square root, \sqrt{X} , and the logarithm, $\log X$, are often used as transformations in this way. The scores made by students in national examinations are frequently rescaled so that they appear to follow a normal curve.

4. With measurement data, many investigations have as their purpose the estimation of averages—the average life of a battery, the average income of plumbers, and so on. Even if the distribution in the original population is far from normal, the distribution of sample averages tends to become normal, under a wide variety of conditions, as the size of sample increases. This is perhaps the single most important reason for the use of the normal.

5. Finally, many results that are useful in statistical work, although strictly true only when the population is normal, hold well enough for rough-and-ready use when samples come from non-normal populations. When presenting such results we shall try to indicate how well they stand up under non-normality.

2.3—Tables of the normal distribution. Since the normal curve depends on the two parameters μ and σ , there are a great many different normal curves. All standard tables of this distribution are for the distribution with $\mu = 0$ and $\sigma = 1$. Consequently if you have a measurement X with mean μ and standard deviation σ and wish to use a table of the normal distribution, you must rescale X so that the mean becomes 0 and the standard deviation becomes 1. The rescaled measurement is given by relation

$$Z = \frac{X - \mu}{\sigma}$$

The quantity Z goes by various names—a *standard normal variate*, a *standard normal deviate*, a *normal variate in standard measure*, or, in education and psychology, a *standard score* (although this term sometimes has a slightly different meaning). To transform back from the Z scale to the X scale, the formula is

$$X = \mu + \sigma Z$$

There are two principal tables.

Table of ordinates. Table A 2 (p. 547) gives the ordinates or heights of the standard normal distribution. The formula for the ordinate is

$$y = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}$$

These ordinates are used when graphing the normal curve. Since the curve is symmetrical about the origin, the heights are presented only for positive values of Z . Here is a worked example.

EXAMPLE 1 - Suppose that we wish to sketch the normal curve for a variate X that has $\mu = 3$ and $\sigma = 1.6$. What is the height of this curve at $X = 2$?

Step 1. Find $Z = (2 - 3)/1.6 = -0.625$.

Step 2. Read the ordinate in table A 2 for $Z = 0.625$. In the table, the Z entries are given to two decimal places only. For $Z = 0.62$ the ordinate is 0.3292 and for $Z = 0.63$ the ordinate is 0.3271. Hence we take 0.328 for $Z = 0.625$.

Step 3. Finally, divide the ordinate 0.328 by σ , getting $0.328/1.6 = 0.205$ as the answer. This step is needed because if you look back at the formula in section 2.1 for the ordinate of the general normal curve, you will see a σ in the denominator that does not appear in the tabulated curve.

Table of the cumulative distribution. Table A 3 (p. 548) is much more frequently used than Table A 2. This gives, for any positive value of Z , the area under the curve from the origin up to the point Z . It shows, for any positive Z , the probability that a variate drawn at random from the standard normal distribution will have a value lying between 0 and Z . The word *cumulative* is used because if we think of the frequency distribution of a very large sample, with many classes, the area under the curve represents the total or cumulative frequency in all classes lying between 0 and Z , divided by the total sample size so as to give a cumulative relative frequency. In the limit, as the sample size increases indefinitely, this becomes the probability that a randomly drawn member lies between 0 and Z .

As a reminder the area tabulated in Table A 3 is shown in figure 2.3.1. Since different people have tabulated different types of area under the normal curve, it is essential, when starting to use any table, to understand clearly what area has been tabulated.

First, a quick look at table A 3. At $Z = 0$ the area is, of course, zero. At $Z = 3.9$, or any larger value, the area is 0.5000 to four decimal places. It follows that the probability of a value of Z lying between -3.9 and

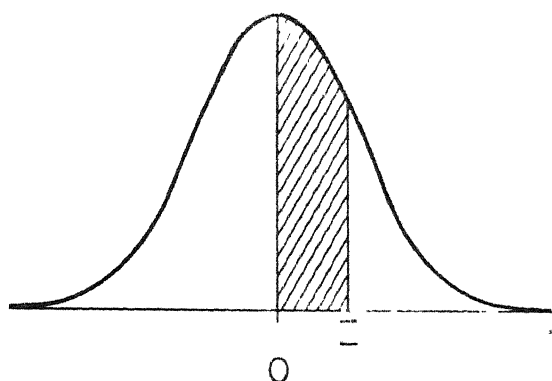


FIG. 2.3.1 The shaded area is the area tabulated in table A 3 for positive values of Z

+3.9 is 1.0000 to four decimals, remembering that the curve is symmetrical about the origin. This means that any value drawn from a standard normal distribution is practically certain to lie between -3.9 and $+3.9$. At $Z = 1.0$, the area is 0.3413. Thus the probability of a value lying between -1 and $+1$ is 0.6826. This verifies a previous remark (section 2.1) that over $2/3$ of the observations in a general normal distribution lie in the interval $\mu \pm \sigma$. Similarly, for $Z = 2$ the area is 0.4772, corresponding to the result that about 95% of the observations (more accurately 95.44%) will lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.

When using table A 3 you will often want probabilities represented by areas different from those tabulated. If A is the area in table A 3, the following table shows how to obtain the probabilities most commonly needed.

TABLE 2.3.1
FORMULAS FOR FINDING PROBABILITIES RELATED TO THE NORMAL DISTRIBUTION

Probability of a Value	Formula
(1) Lying between O and Z	A
(2) Lying between $-Z$ and Z	$2A$
(3) Lying outside the interval $(-Z, Z)$	$1 - 2A$
(4) Less than Z (Z positive)	$0.5 + A$
(5) Less than Z (Z negative)	$0.5 - A$
(6) Greater than Z (Z positive)	$0.5 - A$
(7) Greater than Z (Z negative)	$0.5 + A$

Verification of these formulas is left as an exercise. A few more complex examples will be worked:

EXAMPLE 2—What is the probability that a normal deviate lies between -1.62 and $+0.28$? We have to split the interval into two parts: from -1.62 to 0 , and from 0 to 0.28 . From table A 3, the areas for the two parts are, respectively, 0.4474 and 0.1103, giving 0.5577 as the answer.

38 Chapter 2: Sampling From a Normally Distributed Population

EXAMPLE 3—What is the probability that a normal deviate lies between -2.67 and -0.59 ? In this case we take the area from -2.67 to 0 , namely 0.4972 , and subtract from it the area from -0.59 to 0 , namely 0.2224 , giving 0.2748 .

EXAMPLE 4—The heights of a large sample of men were found to be approximately normally distributed with mean $= 67.56$ inches and standard deviation $= 2.57$ inches. What proportion of the men have heights less than 5 feet 2 inches? We must first find Z .

$$Z = \frac{X - \mu}{\sigma} = \frac{62 - 67.56}{2.57} = -2.163$$

The probability wanted is the probability of a value less than Z , where Z is negative. We use formula (5) in table 2.3.1. Reading table A 3 at $Z = 2.163$, we get $A = 0.4847$, interpolating mentally between $Z = 2.16$ and $Z = 2.17$. From formula (5), the answer is $0.5 - A$, or 0.0153 . About $1\frac{1}{2}\%$ of the men have heights less than 5 ft. 2 in.

EXAMPLE 5—What height is exceeded by 5% of the men? The first step is to find Z we use formula (6) in table 2.3.1, writing $0.5 - A = 0.05$, so that $A = 0.45$. We now look in table A 3 for the value of Z such that $A = 0.45$. The value is $Z = 1.645$. Hence the actual height is

$$X = \mu + \sigma Z = 67.56 + (2.57)(1.645) = 71.79 \text{ inches,}$$

just under 6 feet.

Some examples to be worked by the reader follow:

EXAMPLE 2.3.1—Using table A 2, (i) at the origin, what is the height of a normal curve with $\sigma = 2$? (ii) for any normal curve, at what value of X is the height of the curve one-tenth of the height at the origin? Ans. (i) 0.1994 ; (ii) at the value $X = \mu + 2.15\sigma$.

EXAMPLE 2.3.2—Using table A 3, show that 92.16% of the items in a normally distributed population lie between -1.76σ and $+1.76\sigma$.

EXAMPLE 2.3.3—Show that 65.24% of the items in a normal population lie between $\mu - 1.1\sigma$ and $\mu + 0.8\sigma$.

EXAMPLE 2.3.4—Show that 13.59% of the items lie between $Z = 1$ and $Z = 2$.

EXAMPLE 2.3.5—Show that half the population lies in the interval from $\mu - 0.6745\sigma$ and $\mu + 0.6745\sigma$. The deviation 0.6745σ , formerly much used, is called the *probable error* of X . Ans. You will have to use interpolation. You are seeking a value of Z such that the area from O to Z is 0.2500 . $Z = 0.67$ gives 0.2486 and $Z = 0.68$ gives 0.2517 . Since $0.2500 - 0.2486 = 0.0014$, and $0.2517 - 0.2486 = 0.0031$, we need to go $14/31$ of the distance from 0.67 to 0.68 . Since $14/31 = 0.45$, the interpolate is $Z = 0.6745$.

EXAMPLE 2.3.6—Show that 1% of the population lies outside the limits $Z = \pm 2.575$

EXAMPLE 2.3.7—For the heights of men, with $\mu = 67.56$ inches and $\sigma = 2.57$ inches, what percentage of the population has heights lying between 5 feet 5 inches and 5 feet 10 inches? Compute your Z 's to two decimals only. Ans. 67% .

EXAMPLE 2.3.8—The specification for a manufactured component is that the pressure at a certain point must not exceed 30 pounds. A manufacturer who would like to enter this market finds that he can make components with a mean pressure $\mu = 28$ lbs., but the pressure varies from one specimen to another with a standard deviation $\sigma = 1.6$ lbs. What proportion of his specimens will fail to meet the specification? Ans. 10.6% .

EXAMPLE 2.3.9—By quality control methods it may be possible to reduce σ in the previous example while keeping μ at 28 lbs. If the manufacturer wishes only 2% of his specimens to be rejected, what must his σ be? Ans. 0.98 lbs.

2.4—Estimators of μ and σ . While μ and σ are seldom known, they may be estimated from random samples. To illustrate the estimation of the parameters, we turn to the data reported from a study. In 1936 the Council on Foods of the American Medical Association sampled the vitamin C content of commercially canned tomato juice by analyzing a specimen from each of the 17 brands that displayed the seal of the Council (1). The vitamin C concentrations in mg. per 100 gm. are as follows (slightly altered for easier use):

16, 22, 21, 20, 23, 21, 19, 15, 13, 23, 17, 20, 29, 18, 22, 16, 25

Estimation of μ . Assuming random sampling from a normal population, μ is estimated by an average called the *mean of the sample* or, more briefly, the *sample mean*. This is calculated by the familiar process of dividing the sum of the observations, X , by their number. Representing the sample mean by \bar{X} ,

$$\bar{X} = 340/17 = 20 \text{ mg. per 100 grams of juice}$$

The symbol, \bar{X} is often called “bar- X ” or “ X -bar.” We say that this sample mean is an estimator of μ or that μ is estimated by it.

Estimation of σ . The simplest estimator of σ is based on the *range* of the sample observations, that is, the difference between the largest and smallest measurements. For the vitamin C data,

$$\text{range} = 29 - 13 = 16 \text{ mg./100 gm.}$$

From the range, sigma is estimated by means of a multiplier which depends on the sample size. The multiplier is shown in the column headed “ σ /Range” in table 2.4.1 (2, 3). For $n = 17$, halfway between 16 and 18, the multiplier is 0.279, so that

$$\sigma \text{ is estimated by } (0.279)(16) = 4.46 \text{ mg./100 gm.}$$

Looking at table 2.4.1 you will notice that the multiplier decreases as n becomes larger. This is because the sample range tends to increase as the sample size increases, although the population σ remains unchanged. Clearly if we start with a sample of size 2 and keep adding to it, the range must either stay constant or go up with each addition.

Quite easily, then, we have made a *point estimate* of each parameter of a normal population; these estimators constitute a summary of the information contained in the sample. The sample mean cannot be improved upon as an estimate of μ , but we shall learn to estimate σ more efficiently. Also we shall learn about interval estimates and tests of hypotheses. Before doing so, it is worthwhile to examine our sample in greater detail.

The first point to be clarified is this: What population was represented by the sample of 17 determinations of vitamin C? We raised this question tardily; it is the first one to be considered in analyzing any sampling. The report makes it clear that not all brands were sampled, only the seventeen allowed to display the seal of the Council. The dates of the

TABLE 2.4.1
 RATIO OF σ TO RANGE IN SAMPLES OF n FROM THE NORMAL DISTRIBUTION. EFFICIENCY
 OF RANGE AS ESTIMATOR OF σ . NUMBER OF OBSERVATIONS WITH
 RANGE TO EQUAL 100 WITH s

n	$\frac{\sigma}{\text{Range}}$	Relative Efficiency	Number per 100	n	$\frac{\sigma}{\text{Range}}$	Relative Efficiency	Number per 100
2	0.886	1.000	100	12	0.307	0.815	123
3	.591	.992	101	14	.204	.783	128
4	.486	.975	103	16	.283	.753	133
5	.430	.955	105	18	.275	.726	138
6	.395	.933	107	20	.268	.700	143
7	.370	.912	110	30	.245	.604	166
8	.351	.890	112	40	.231	.536	186
9	.337	.869	115	50	.222	.49	204
10	.325	.850	118				

packs were mostly August and September of 1936, about a year before the analyses were made. The council report states that the vitamin concentration "may be expected to vary according to the variety of the fruit, the conditions under which the crop has been grown, the degree of ripeness and other factors." About all that can be said, then, is that the sampled population consisted of those year-old containers still available to the 17 selected packers.

2.5—The array and its graphical representation. Some of the more intimate features of a sample are shown by arranging the observations in order of size, from low to high, in an *array*. The array of vitamin contents is like this:

13, 15, 16, 16, 17, 18, 19, 20, 20, 21, 21, 22, 22, 23, 23, 25, 29

For a small sample the array serves some of the same purposes as the frequency distribution of a large one.

The range, from 13 to 29, is now obvious. Also, attention is attracted to the concentration of the measures near the center of the array and to their thinning out at the extremes. In this way the sample may reflect the distribution of the normal population from which it was drawn. But the smaller the sample, the more erratic its reflection may be.

In looking through the vitamin C contents of the several brands, one is struck by their variation. What are the causes of this variation? Different processes of manufacture, perhaps, and different sources of the fruit. Doubtless, also, the specimens examined, being themselves samples of their brands, differed from the brand means. Finally, the laboratory technique of evaluation is never perfectly accurate. Variation is the essence of statistical data.

Figure 2.5.1 is a graphical representation of the foregoing array of 17 vitamin determinations. A dot represents each item. The distance of the

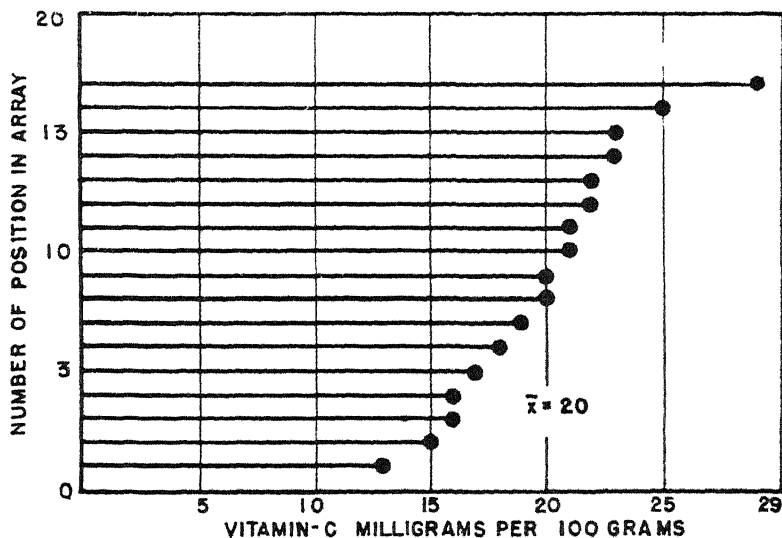


FIG. 2.5.1— Graphical representation of an array. Vitamin C data.

dot from the vertical line at the left, proportional to the concentration of ascorbic acid in a brand specimen, is read in milligrams per 100 grams on the horizontal scale.

The diagram brings out vividly not only the variation and the concentration in the sample, but also two other characteristics: (i) the rather symmetrical occurrence of the values above and below the mean, and (ii) the scarcity of both extremely small and extremely large vitamin C contents, the bulk of the items being near the middle of the set. These features recur with notable persistence in samples from normal distributions. For many variables associated with living organisms there are averages and ranges peculiar to each, reflecting the manner in which each seems to express itself most successfully. These norms persist despite the fact that individuals enjoy a considerable freedom in development. A large part of our thinking is built around ideas corresponding to such statistics. Each of the words, *pie*, *daisy*, *man*, raises an image which is quantitatively described by summary numbers. It is difficult to conceive of progress in thought until memories of individuals are collected into concepts like averages and ranges of distributions.

2.6—Algebraic notation. The items in any set may be represented by

$$X_1, X_2, X_3, \dots X_n,$$

where the subscripts 1, 2, . . . n , may specify position in the set of n items (not necessarily an array). The three dots accompanying these symbols

42 Chapter 2: Sampling From a Normally Distributed Population

are read “and so on.” Matching the symbols with the values in section 2.4,

$$X_1 = 16, X_2 = 22, \dots X_{17} = 25 \text{ mg./100 gm.}$$

The sample mean is represented by \bar{X} , so that

$$\bar{X} = (X_1 + X_2 + \dots X_n)/n$$

This is condensed into the form,

$$\bar{X} = (\Sigma X)/n$$

where X stands for every item successively. The symbol, ΣX , is read “summation X ” or “sum of the X .” Applying this formula to the vitamin C concentrations,

$$\Sigma X = 340, \text{ and } \bar{X} = 340/17 = 20 \text{ mg./100 gm.}$$

2.7—Deviations from sample mean. The individual variations of the items in a set of data may be well expressed by the *deviations* of these items from some centrally located number such as the sample mean. For example, the deviation-from-mean of the first X -value is

$$16 - 20 = -4 \text{ mg. per 100 gm.}$$

That is, this specimen falls short of \bar{X} by 4 mg./100 gm. Of special interest is the whole set of deviations calculated from the array in section 2.5:

$$-7, -5, -4, -4, -3, -2, -1, 0, 0, 1, 1, 2, 2, 3, 3, 5, 9$$

These deviations are represented graphically in figure 2.5.1 by the distances of the dots from the vertical line drawn through the sample mean.

Deviations are almost as fundamental in our thinking as are averages. “What a whale of a pig” is a metaphor expressing astonishment at the deviation of an individual’s size from the speaker’s concept of the normal. Gossip and news are concerned chiefly with deviations from accepted standards of behavior. Curiously, interest is apt to center in departures from norm, rather than in that background of averages against which the departures achieve prominence. Statistically, freaks are freaks only because of their large deviations.

Deviations are represented symbolically by lower case letters. That is:

$$x_1 = X_1 - \bar{X}$$

$$x_2 = X_2 - \bar{X}$$

$$\cdot \quad \cdot \quad \cdot$$

$$\cdot \quad \cdot \quad \cdot$$

$$x_n = X_n - \bar{X}$$

Just as X may represent any of the items in a set, or all of them in succession, so x represents deviations from sample mean. In general,

$$x = X - \bar{X}$$

It is easy to prove the algebraic result that the sum of a set of deviations from the mean is zero; that is, $\Sigma x = 0$. Look at the set of deviations $x_1 = X_1 - \bar{X}$, and so on (foot of p. 42). Instead of adding the column of values x_i we can obtain the same result by adding the column of values X_i and subtracting the sum of the column of values \bar{X} . The sum of the column of values X_i is the expression ΣX . Further, since there are n items in a column, the sum of the column of values \bar{X} is just $n\bar{X}$. Thus we have the result

$$\Sigma x = \Sigma X - n\bar{X}$$

But the mean $\bar{X} = \Sigma X/n$, so that $n\bar{X} = \Sigma X$, and the right-hand side is zero. It follows from this theorem that the *mean* of the deviations is also zero.

This result is useful in proving several standard statistical formulas. When it is applied to a specific sample of data, there is a slight snag. If the sample mean \bar{X} does not come out exactly, we have to round it. As a result of this rounding, the numerical sum of the deviations will not be exactly zero. Consider a sample with the values 1, 7, 8. The mean is $16/3$, which we might round to 5.3. The deviations are then -4.3 , $+1.7$, and $+2.7$, adding to $+0.1$. Thus in practice the sum of the deviations is zero, apart from rounding errors.

EXAMPLE 2.7.1—The weights of 12 staminate hemp plants in early April at College Station, Texas (9), were approximately:

13, 11, 16, 5, 3, 18, 9, 9, 8, 6, 27, and 7 grams

Array the weights and represent them graphically. Calculate the sample mean, 11 grams, and the deviations therefrom. Verify the fact that $\Sigma x = 0$. Show that σ is estimated by 7.4 grams.

EXAMPLE 2.7.2—The heights of 11 men are 64, 70, 65, 69, 68, 67, 68, 67, 66, 72 and 61 inches. Compute the sample mean and verify it by summing the deviations. Are the numbers of positive and negative deviations equal, or only their sums?

EXAMPLE 2.7.3—The weights of 11 forty-year-old men were 148, 154, 158, 160, 161, 162, 166, 170, 182, 195, and 236 pounds. Notice the fact that only three of the weights exceed the sample mean. Would you expect weights of men to be normally distributed?

EXAMPLE 2.7.4—In a sample of 48 observations you are told that the standard deviation has been computed and is 4.0 units. Glancing through the data, you notice that the lowest observation is 39 and the highest 76. Does the reported standard deviation look reasonable?

EXAMPLE 2.7.5 Ten patients troubled with sleeplessness each received a nightly dose of a sedative for one period, while in another period they received no sedative (4). The average hours of sleep per night for each patient during each two-week period are as follows:

Patient	1	2	3	4	5	6	7	8	9	10
Sedative	1.3	1.1	6.2	3.6	4.9	1.4	6.6	4.5	4.3	6.1
None	0.6	1.1	2.5	2.8	2.9	3.0	3.2	4.7	5.5	6.2

Calculate the 10 differences (Sedative - None). Might these differences be a sample from a normal population of differences? How would you describe this population? (You might want to ask for more information.) Assuming that the differences are normally distributed, estimate μ and σ for the population of differences. Ans. +0.75 hours and 1.72 hours

EXAMPLE 2.7.6—If you have two sets of data that are paired as in the preceding example, and if you have calculated the resulting set of differences, prove algebraically that the sample mean of the differences is equal to the difference between the sample means of the two sets. Verify this result for the data in example 2.7.5.

2.8—Another estimator of σ ; the sample standard deviation. The range, dependent as it is on only the two extremes in a sample, usually has a more variable sampling distribution than an estimator based on the whole set of deviations-from-mean in a sample, not just the largest and smallest. What kind of average is appropriate to summarize these deviations, and to estimate σ with the least sampling variation?

Clearly, the sample mean of the deviations is useless as an estimator because it is always zero. But a natural suggestion is to ignore the signs, calculating the sample mean of the absolute values of the deviations. The resulting measure of variation, the *mean absolute deviation*, had a considerable vogue in times past. Now, however, we use another estimator, more efficient and more flexible.

The sample standard deviation. This estimator, denoted by s , is the most widely used in statistical work. The formula defining s is

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} = \sqrt{\frac{\sum x^2}{n - 1}}$$

First, each deviation is squared. Next, the sum of squares, $\sum x^2$, is divided by $(n - 1)$, one less than the sample size. The result is the *mean square* or *sample variance*, s^2 . Finally, the extraction of the square root recovers the original scale of measurement. For the vitamin C concentrations, the calculations are set out in the right-hand part of table 2.8.1. Since the sum of squares of the deviations is 254 and n is 17, we have

$$s^2 = 254/16 = 15.88$$

$$s = \sqrt{15.88} = 3.98 \text{ mg. } 100 \text{ gm}$$

Before further discussion of s is given, its calculation should be fixed in mind by working a couple of examples. Table A.18 is a table of square roots. Hints on finding square roots are given on p. 541.

TABLE 2.8.1
CALCULATION OF THE SAMPLE STANDARD DEVIATION

Observation Number	Vitamin C Concentration Mg Per 100 gm	Deviation From Mean		Deviation Squared
i	X	$x = X - \bar{X}$	\bar{X}	x^2
1	16	-4		16
2	22		2	4
3	21		1	1
4	20	0		0
5	23		+3	9
6	21		+1	1
7	19	-1		1
8	15	-5		25
9	13	-7		49
10	23		+3	9
11	17	-3		9
12	20	0		0
13	29		+9	81
14	18	-2		4
15	22		+2	4
16	16	-4		16
17	25		+5	25
Totals	340	-26	+26	254

EXAMPLE 2.8.1—In five patients with pneumonia, treated with sodium penicillin G, the numbers of days required to bring the temperature down to normal were 1, 4, 5, 7, 3. Compute s for these data and compare it with the estimate based on the range. Ans. $s = 2.24$ days. Range estimate = 2.58 days.

EXAMPLE 2.8.2—Calculate s for the hemp plant weights in example 2.7.1. Ans. 6.7 grams. Compare with your first estimate of σ .

The appearance of the divisor $(n - 1)$ instead of n in computing s^2 and s is puzzling at first sight. The reason cannot be explained fully at this stage, being related to the computation of s from data of more complex structure. The quantity $(n - 1)$ is called the *number of degrees of freedom* in s . Later in the book we shall meet situations in which the number of degrees of freedom is neither n nor $(n - 1)$, but some other quantity. If the practice of using the degrees of freedom as divisor is followed, there is the considerable advantage that the same statistical tables, needed in important applications, serve for a wide variety of types of data.

Division by $(n - 1)$ has one standard property that is often cited. If random samples are drawn from *any* indefinitely large population (not just a normally distributed one) that has a finite value of σ , then the average value of s^2 , taken over all random samples, is exactly equal to σ^2 . Any estimate whose average value over all possible random samples is equal to the population parameter being estimated is called *unbiased*. Thus,

s^2 is an *unbiased estimate* of σ^2 . This property, which says that *on the average* the estimate gives the correct answer, seems a desirable one for an estimate to possess. The property, however, is not as fundamental as one might think, because s is not an unbiased estimate of σ . If we want s to be an unbiased estimate of σ in normal populations, we must use a divisor that is neither $(n - 1)$ nor n .

2.9—Comparison of the two estimators of σ . You now have two estimators of σ , one of them easier to calculate than the other, but less efficient. You need to know what is meant by “less efficient” and what governs the choice of estimate. Suppose that we draw a large number of random samples of size 10 from a normal population. For each sample we can compute the estimate of σ obtained from the range, and the estimate s . Thus we can form two frequency-distributions, one showing the distribution of the range estimate, the other showing the distribution of s . The distribution of s is found to be more closely grouped about σ ; that is, s usually gives a more accurate estimate of σ . Going a step further, it can be shown that the range estimate, computed from normal samples of size 12, has roughly the same frequency distribution as that of s in samples of size 10. We say that in samples of size 10 the *relative efficiency* of the range estimator to s is about $10/12$, or more accurately 0.850. The relative efficiencies and the relative sample sizes appear in the third and fourth columns of table 2.4.1 (p. 40). In making a choice we have to weigh the cost of more observations. If observations are costly, it is cheaper to compute s .

Actually, both estimators are extensively used. Note that the relative efficiency of the range estimator remains high up to samples of sizes 8 to 10. In many operations, σ is estimated in practice by combining the estimates from a substantial number of small samples. For instance, in controlling the quality of an industrial process, small samples of the manufactured product are taken out and tested frequently, say every 15 minutes or every hour. Samples of size 5 are often used, the range estimator being computed from each sample and plotted on a time-chart. The efficiency of a single range estimate in a sample of size 5 is 0.955, and the average of a series of ranges has the same efficiency.

The estimate from the range is an easy approximate check on the computation of s . In these days, electronic computing machines are used more and more for routine computations. Unless the investigator has learned how to program, one consequence is that the details of his computations are taken out of his hands. Errors in making the programmers understand what is wanted and errors in giving instructions to the machines are common. There is therefore an increasing need for quick approximate checks on all the standard statistical computations, which the investigator can apply when his results are handed to him. If a table of σ/Range is not at hand, two rough rules may help. For samples up to size 10, divide the range by \sqrt{n} to estimate σ . Remember also:

If n is near this number	Then σ is roughly estimated by dividing range by
5	2
10	3
25	4
100	5

The range estimator and s are both sensitive to gross errors, because a gross error is likely to produce a highest or lowest sample member that is entirely false.

EXAMPLE 2.9.1—In a sample of size 2, with measurements X_1 and X_2 , show that s is $|X_1 - X_2|/\sqrt{2} = 0.707|X_1 - X_2|$, and that the range estimator is $0.886|X_1 - X_2|$, where the vertical lines denote the *absolute value*. The reason for the different multipliers is that the range estimator is constructed to be an unbiased estimator of σ , while s is not, as already mentioned.

EXAMPLE 2.9.2—The birth weights of 20 guinea pigs were: 30, 30, 26, 32, 30, 23, 29, 31, 36, 30, 25, 34, 32, 24, 28, 27, 38, 31, 34, 30 grams. Estimate σ in 3 ways: (i) by the rough approximation, one-fourth of the range (Ans. 3.8 gm.); (ii) by use of the fraction, 0.268, in table 2.4.1 (Ans. 4.0 gm.); (iii) by calculating s (Ans. 3.85 gm.). N.B.: Observe the time required to calculate s .

EXAMPLE 2.9.3—In the preceding example, how many birth weights would be required to yield the same precision if the range were used instead of s ? Ans. about 29 weights.

EXAMPLE 2.9.4—Suppose you lined up according to height 16 freshmen, then measured the height of the shortest, 64 inches, and the tallest, 72 inches. Would you accept the midpoint of the range, $(64 + 72)/2 = 68$ inches as a rough estimate of μ , and $8/3 = 2.7$ inches as a quick-and-easy estimate of σ ?

EXAMPLE 2.9.5—In a sample of 3 the values are, increasing order, X_1 , X_2 , and X_3 . The range estimate of σ is $0.591(X_3 - X_1)$. If you are ingenious at algebra, show that s always lies between $(X_3 - X_1)/2 = 0.5(X_3 - X_1)$, and $(X_3 - X_1)/\sqrt{3} = 0.578(X_3 - X_1)$. Verify the two extreme cases from the samples 0, 3, 6, in which $s = 0.5(X_3 - X_1)$ and 0, 0, 6, in which $s = 0.578(X_3 - X_1)$.

2.10—Hints on the computation of s . Two results in algebra help to shorten the calculation of s . Both give quicker ways of finding $\sum x^2$. If G is any number, there is an algebraic identity to the effect that

$$\sum x^2 = \sum (X - \bar{X})^2 = \sum (X - G)^2 - (\sum X - nG)^2/n$$

An equivalent alternative form is

$$\sum x^2 = \sum (X - \bar{X})^2 = \sum (X - G)^2 - n(\bar{X} - G)^2$$

These expressions are useful when s has to be computed without the aid of a calculating machine (a task probably confined mainly to students nowadays). Suppose the sample total is $\sum X = 350$ and $n = 17$. The mean \bar{X} is $350/17 = 20.59$. If the X 's are whole numbers, it is troublesome to take deviations from a number like 20.59, and still more so to square the numbers without a machine. The trick is to take G (sometimes called the

guessed or working mean) equal to 20. Find the deviations of the X 's from 20 and the sum of squares of these deviations, $\Sigma(X - G)^2$. To get Σx^2 , you have only to subtract n times the square of the difference between \bar{X} and G , or, in this case, $17(0.59)^2 = 5.92$.

Proof of the identity. We shall denote a typical value in the sample by X_i , where the subscript i goes from 1 to n . Write

$$X_i - G = (X_i - \bar{X}) + (\bar{X} - G)$$

Squaring both sides, we have

$$(X_i - G)^2 = (X_i - \bar{X})^2 + 2(X_i - \bar{X})(\bar{X} - G) + (\bar{X} - G)^2$$

We now add over the n members of the sample. In the middle term on the right, the term $2(\bar{X} - G)$ is a constant multiplier throughout this addition, since this term does not contain the subscript i that changes from one member of the sample to another. Hence

$$\Sigma 2(X_i - \bar{X})(\bar{X} - G) = 2(\bar{X} - G)\Sigma(X_i - \bar{X}) = 0,$$

since, as we have seen previously, the sum of the deviations from the sample mean is always zero. This gives

$$\Sigma(X_i - G)^2 = \Sigma(X_i - \bar{X})^2 + n(\bar{X} - G)^2$$

noting that the sum of the constant term $(\bar{X} - G)^2$ over the sample is $n(\bar{X} - G)^2$. Moving this term to the other side, we get

$$\Sigma(X_i - G)^2 - n(\bar{X} - G)^2 = \Sigma(X_i - \bar{X})^2$$

This completes the proof.

Incidentally, the result shows that for any value of G , $\Sigma(X_i - \bar{X})^2$ is always *smaller* than $\Sigma(X_i - G)^2$, unless $G = \bar{X}$. The sample mean has the property that the sum of squares of deviations from it is a minimum.

The second algebraic result, a particular case of the first, is used when a calculating machine is available. Put $G = 0$ in the first result in this section. We get

$$\Sigma x^2 = \Sigma(X - \bar{X})^2 = \Sigma X^2 - (\Sigma X)^2/n$$

This result enables us to find Σx^2 without computing any of the deviations. For a set of positive numbers X_i , most calculating machines will compute the sum of squares, ΣX^2 , and the sum, ΣX , simultaneously, without writing down any intermediate figures. To get Σx^2 , we square the sum, dividing by n , to give $(\Sigma X)^2/n$, and subtract this from the original sum of squares, ΣX^2 . The computation will be illustrated for the 17 vitamin C concentrations. Earlier, as mentioned, these data were altered slightly to simplify the presentation. The actual determinations were as follows.

16, 22, 21, 20, 23, 22, 17, 15, 13, 22, 17, 18, 29, 17, 22, 16, 23

The only figures that need be written down are shown in table 2.10.1.

TABLE 2.10.1
COMPUTING THE SAMPLE MEAN AND SUM OF SQUARES OF DEVIATIONS
WITH A CALCULATING MACHINE

$n = 17$	$\Sigma X^2 = 6,773$
$\Sigma X = 333$	$(\Sigma X)^2/n = 6,522.88$
$\bar{X} = 19.6 \text{ mg. per } 100 \text{ gm.}$	$\Sigma x^2 = 250.12$
$s^2 = 250.12/16 = 15.63$	
$s = \sqrt{15.63} = 3.95$	

When using this method, remember that any constant number can be subtracted from all the X_i without changing s . Thus if your data are numbers like 1032, 1017, 1005, and so on, they can be read as 32, 17, 5, and so on, when following the method in table 2.10.1

EXAMPLE 2.10.1—For those who need practice in using a guessed mean, here is a set of numbers for easy computation:

15, 12, 10, 10, 10, 8, 7, 7, 4, 4, 1

First calculate $\bar{X} = 8$ and $s = 4$ by finding deviations from the sample mean. Then try various guessed means, such as 5, 10, and 1. Continue until you convince yourself that the answers, $\bar{X} = 8$ and $s = 4$, can be reached regardless of the value chosen for G . Finally, try $G = 0$. Note: With a guessed mean, \bar{X} can be found without having to add the X_i , by the relation

$$\bar{X} = G + [\Sigma(X - G)]/n$$

where the quantity $\Sigma(X - G)$ is the sum of your deviations from the guessed mean G

EXAMPLE 2.10.2—For the ten patients in a previous example, the average differences in hours of sleep per night between sedative and no sedative were (in hours) 0.7, 0.0, 3.7, 0.8, 2.0, -1.6, 3.4, -0.2, -1.2, -0.1. With a calculating machine, compute s by the short-cut method in table 2.10.1. Ans. $s = 1.79$ hrs. The range method gave 1.72 hrs

EXAMPLE 2.10.3—Without finding deviations from \bar{X} and without using a calculating machine, compute Σx^2 for the following measurements: 961, 953, 970, 958, 950, 951, 957. Ans. 286.9.

2.11—The standard deviation of sample means. With measurement data, as mentioned previously, the purpose of an investigation is often to estimate an average or total over a population (average selling price of houses in a town, total wheat crop in a region). If the data are a random sample from a population, the sample mean \bar{X} is used to estimate the corresponding average over the population. Further, if the number of items N in the population is known, the quantity $N\bar{X}$ is an estimator of the population total of the X 's. This brings up the question: How accurate is a sample mean as an estimator of the population mean?

As usual, a question of this type can be examined either experimentally or mathematically. With the experimental approach, we first find or construct a population that seems typical of the type of population encountered in our work. Suppose that we are particularly interested in

samples of size 100. We draw a large number of random samples of size 100, computing the sample mean \bar{X} for each sample. In this way we form a frequency distribution of the sample means, or graph the frequencies in a histogram. Since the mean of the population is known, we can find out how often the sample mean is satisfactorily close to the population mean, and how often it gives a poor estimate.

Much mathematical work has been done on this problem and it has produced two of the most exciting and useful results in the whole of statistical theory. These results, which are part of every statistician's stock in trade, will be stated first. Some experimental verification will then be presented for illustration. The first result gives the mean and standard deviation of \bar{X} in repeated sampling; the second gives the shape of the frequency distribution of \bar{X} .

Mean and standard deviation of \bar{X} . If repeated random samples of size n are drawn from any population (not necessarily normal) that has mean μ and standard deviation σ , the frequency distribution of the sample means \bar{X} in these repeated samples has mean μ and standard deviation σ/\sqrt{n} .

This result says that under random sampling the sample mean \bar{X} is an unbiased estimator of μ : on the average, in repeated sampling, it will be neither too high nor too low. Further, the sample means have less variation about μ than the original observations. The larger the sample size, the smaller this variation becomes.

Students sometimes find it difficult to reach the point at which the phrase "the standard deviation of \bar{X} " has a concrete meaning for them. Having been introduced to the idea of a standard deviation, it is not too hard to feel at home with a phrase like "the standard deviation of a man's height," because every day we see tall men and short men, and realize that this standard deviation is a measure of the extent to which heights vary from one man to another. But usually when we have a sample, we calculate a *single* mean. Where does the variation come from? It is the variation that would arise if we drew repeated samples from the population that we are studying and computed the mean of each sample. The experimental samplings presented in this chapter and in chapter 3 may make this concept more realistic.

The standard deviation of \bar{X} , σ/\sqrt{n} , is often called, alternatively, the *standard error of \bar{X}* . The terms "standard deviation" and "standard error" are synonymous. When we are studying the frequency distribution of an estimator like \bar{X} , its standard deviation supplies information about the amount of error in \bar{X} when used to estimate μ . Hence, the term "standard error" is rather natural. Normally, we would not speak of the standard error of a man's height, because if a man is unusually tall, this does not imply that he has made a mistake in his height.

The quantity $N\bar{X}$, often used to estimate a total over the population, is also an unbiased estimator under random sampling. Since N is simply a fixed number, the mean of $N\bar{X}$ in repeated sampling is $N\mu$, which, by the definition of μ , is the correct population total. The standard error of

$N\bar{X}$ is $N\sigma/\sqrt{n}$. Another frequently used result is that the sample total, $\Sigma X = n\bar{X}$, has a standard deviation $n\sigma/\sqrt{n}$, or $\sigma\sqrt{n}$.

2.12—The frequency distribution of sample means. The second major result from statistical theory is that, whatever the shape of the frequency distribution of the original population of X 's, the frequency distribution of \bar{X} in repeated random samples of size n tends to become normal as n increases. To put the result more specifically, recall that if we wish to express a variable X in *standard measure*, so that its mean is zero and its standard deviation is 1, we change the variable from X to $(X - \mu)/\sigma$. For \bar{X} , the corresponding expression in standard measure (*sm*) is

$$\bar{X}_{sm} = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}$$

As n increases, the probability that \bar{X}_{sm} lies between any two limits L_1 and L_2 becomes more and more equal to the probability that the standard normal deviate Z lies between L_1 and L_2 . By expressing \bar{X} in standard measure, table A 3 (the cumulative normal distribution) can be used to approximate the probability that \bar{X} itself lies between any two limits. This result, known as the *Central Limit Theorem* (5), explains why the normal distribution and results derived from it are so commonly used with sample means, even when the original population is not normal. Apart from the condition of random sampling, the theorem requires very few assumptions: it is sufficient that σ is finite and that the sample is a random sample from the population.

To the practical worker, a key question is: how large must n be in order to use the normal distribution for \bar{X} ? Unfortunately, no simple general answer is available. With variates like the heights of men, the original distribution is near enough normal so that normality may be assumed for most purposes. In this case a sample with $n = 1$ is large enough. There are also populations, at first sight quite different from the normal, in which $n = 4$ or 5 will do. At the other extreme, some populations require sample sizes well over 100 before the distribution of \bar{X} becomes at all near to the normal distribution.

As illustrations of the Central Limit Theorem, the results of two sampling experiments will be presented. In the first, the population is the population of random digits 0, 1, 2, . . . 9 which we met in chapter 1. This is a discrete population. The variable X has ten possible values 0, 1, 2, . . . 9, and has an equal probability 0.1 of taking any of these values. The frequency distribution of X is represented in the upper part of figure 2.12.1. Clearly, the distribution does not look like a normal distribution. Distributions of this type are sometimes called *uniform*, since every value is equally likely.

Four hundred random samples of size 5 were drawn from the table of random digits (p. 543), each sample being a group of five consecutive

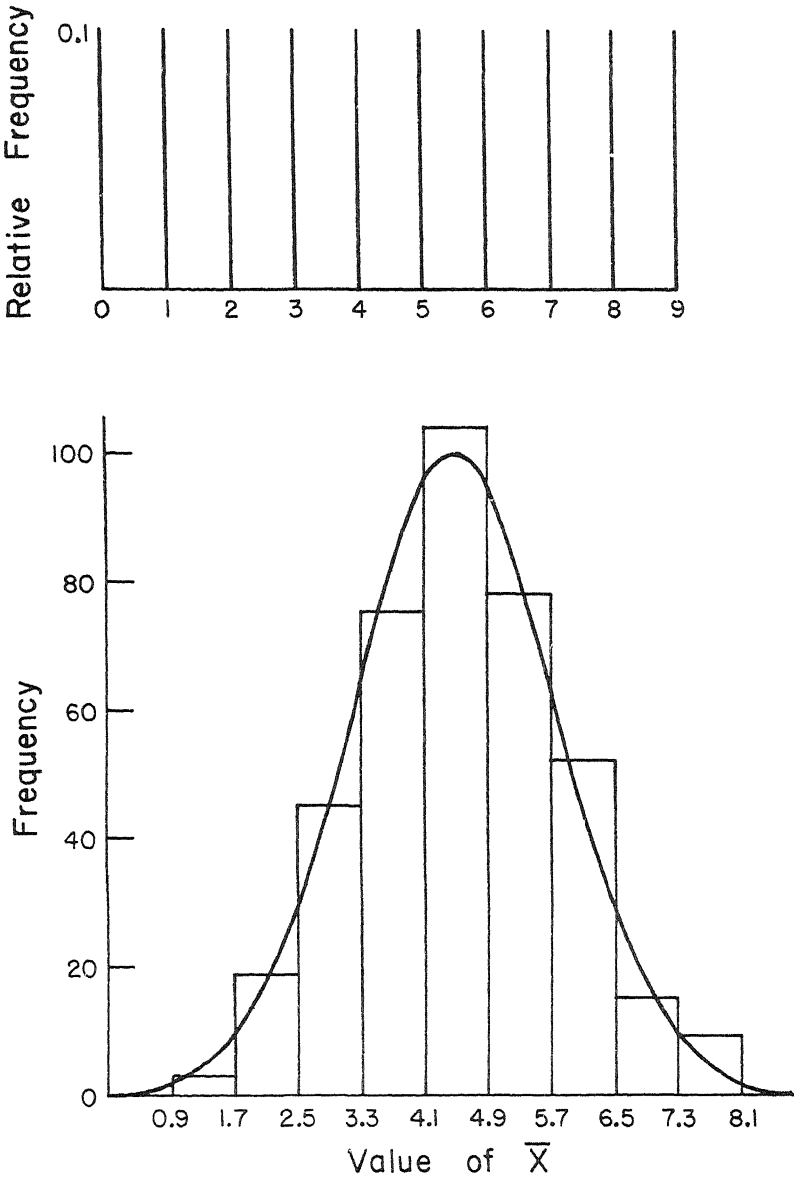


FIG. 2.12.1—Upper part: Theoretical probability distribution of the random digits from 0 to 9. Lower part: Histogram showing the distribution of 400 means of samples of size n drawn from the random digits. The curve is the normal distribution with mean $\mu = 4.5$ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.872}{\sqrt{5}} = 1.284$.

numbers in a column. The frequency distribution of the sample means appears in the lower half of figure 2.12.1. A normal distribution with mean μ and standard deviation $\sigma/\sqrt{5}$ is also shown. The agreement is surprisingly good, considering that the samples are only of size 5.

Calculation of μ and σ . In fitting this normal distribution, the quantities μ and σ were the mean and standard deviation of the original population of random digits. Although the calculation of \bar{X} and s for a sample has been discussed, we have not explained how to calculate μ and σ for a population. In a discrete population, denote the distinct values of the measurement X by X_1, X_2, \dots, X_k . In the population of random digits, $k = 10$, and each value has an equal probability, one-tenth. In a more general discrete population, the value X_i may appear with probability or relative frequency P_i . We could, for example, have a population of digits in which a 0 is 20 times as frequent as a 1. Since the probabilities must add to 1, we have

$$\sum_{i=1}^k P_i = 1$$

The expression on the left is read "the sum of the P_i from i equals 1 to k ."

The population mean μ is defined as

$$\mu = \sum_{i=1}^k P_i X_i$$

Like \bar{X} in a sample, the quantity μ is the average or mean of the values of X_i in the population, noting, however, that each X_i is weighted by its relative frequency of occurrence.

For the random digits, every $P_i = 0.1$. Thus

$$\mu = (0.1)(0 + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9) = (0.1)(45) = 4.5$$

The population σ comes from the deviations $X_i - \mu$. With the random digits, the first deviation is $0 - 4.5 = -4.5$, and the successive deviations are $-3.5, -2.5, -1.5, -0.5, +0.5, +1.5, +2.5, +3.5$, and $+4.5$. The population variance, σ^2 , is defined as

$$\sigma^2 = \sum_{i=1}^k P_i (X_i - \mu)^2$$

Thus, σ^2 is the weighted average of the squared deviations of the values in the population from the population mean. Numerically,

$$\sigma^2 = (0.2)(4.5)^2 + (3.5)^2 + (2.5)^2 + (1.5)^2 + (0.5)^2 = 8.25$$

This gives $\sigma = \sqrt{8.25} = 2.872$; so that $\sigma/\sqrt{5} = 1.284$

There is a shortcut method of finding σ^2 without computing any

deviations: it is similar to the corresponding shortcut formula for Σx^2 . The formula is:

$$\sigma^2 = \sum_{i=1}^k P_i X_i^2 - \mu^2$$

With the *normal* distribution, μ is, as above, the average of the values of X , and σ^2 is the average of the squared deviations from the population mean. Since the normal population is continuous, having an infinite number of values, formulas from the integral calculus are necessary in writing down these definitions.

As a student or classroom exercise, drawing samples of size 5 from the random digit tables is recommended as an easy way of seeing the Central Limit Theorem at work. The total of each sample is quickly obtained mentally. To avoid divisions by 5, work with sample totals instead of means. The sample total, $5\bar{X}$, has mean $(5)(4.5) = 22.5$ and standard deviation $(5)(1.284) = 6.420$ in repeated sampling. In forming the frequency distribution, put the totals 20, 21, 22, 23 in the central class, each class containing four consecutive totals. Although rather broad, this grouping is adequate unless, say, 500 samples have been drawn.

The second sampling experiment illustrates the case in which a large sample size must be drawn if \bar{X} is to be nearly normal. This happens with populations that are markedly skew, particularly if there are a few values very far from the mean. The population chosen consisted of the sizes (number of inhabitants) of U.S. cities having over 50,000 inhabitants in 1950 (6), excluding the four largest cities. All except one have sizes ranging between 50,000 and 1,000,000. The exception, the largest city in the population, contained 1,850,000 inhabitants. The frequency distribution is shown at the top of figure 2.12.2. Note how asymmetrical the distribution is, the smallest class having much the highest frequency. The city with 1,850,000 inhabitants is not shown on this histogram: it would appear about 4 inches to the right of the largest class.

A set of 500 random samples with $n = 25$ and another set with $n = 100$ were drawn. The frequency distributions of the sample means appear in the middle and lower parts of figure 2.12.2. With $n = 25$, the distribution has moved towards the normal shape but is still noticeably asymmetrical. There is some further improvement towards symmetry with $n = 100$, but a normal curve would still be a poor fit. Evidently, samples of 400–500 would be necessary to use the normal approximation with any assurance. Part of the trouble is caused by the 1,850,000 city: the means for $n = 100$ would be more nearly normal if this city had been excluded from the population. On the other hand, the situation would be worse if the four largest cities had been included.

Combining the theorems in this and the previous section, we now have the very useful result that in samples of reasonable size, \bar{X} is approximately normally distributed about μ , with standard deviation or standard error σ/\sqrt{n} .

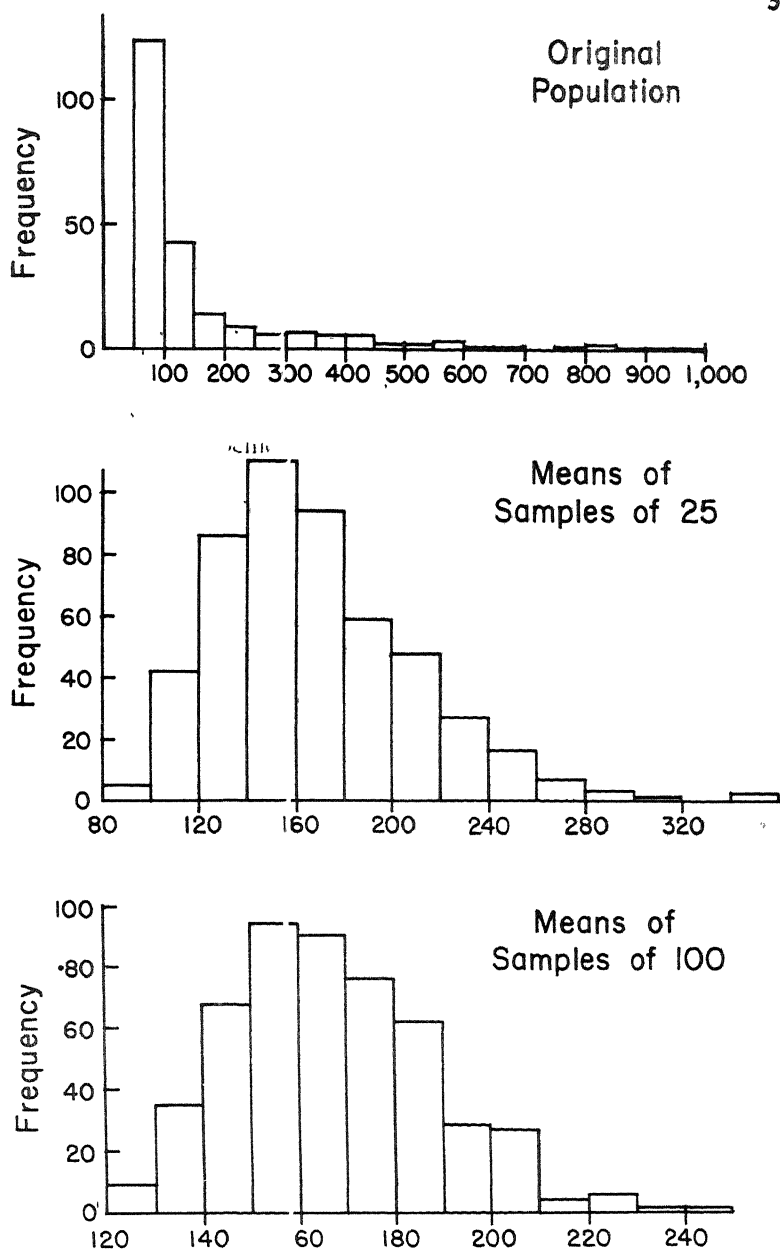


FIG. 2 12.2.—Top part: Frequency distribution of the populations of 228 U.S. cities having populations over 50,000 in 1950. Middle part: Frequency distribution of the means of 500 random samples of size 25. Bottom part: Frequency distribution of the means of 500 random samples of size 100.

56 Chapter 2: Sampling From a Normally Distributed Population

EXAMPLE 2.12.1—A population of heights of men has a standard deviation $\sigma = 2.6$ inches. What is the standard error of the mean of a random sample of (i) 25 men, (ii) 100 men? Ans. (i) 0.52 in. (ii) 0.26 in.

EXAMPLE 2.12.2—In order to estimate the total weight of a batch of 196 bags that are to be shipped, each of a random sample of 36 bags is weighed, giving $\bar{X} = 40$ lbs. Assuming $\sigma = 3$ lbs., estimate the total weight of the 196 bags and give the standard error of your estimate. Ans. 7,840 lbs.; standard error, 98 lbs.

EXAMPLE 2.12.3—In estimating the mean height of a large group of boys with $\sigma = 1.5$ in., how large a sample must be taken if the standard error of the mean height is to be 0.2 in.? Ans. 56 boys.

EXAMPLE 2.12.4—If perfect dice are thrown repeatedly, the probability is $1/6$ that each of the faces 1, 2, 3, 4, 5, 6 turns up. Compute μ and σ for this population. Ans. $\mu = 3.5$, $\sigma = 1.71$.

EXAMPLE 2.12.5—If boys and girls are equally likely, the probabilities that a family of size two contains 0, 1, 2 boys are, respectively, $1/4$, $1/2$, and $1/4$. Find μ and σ for this population. Ans. $\mu = 1$, $\sigma = 1/\sqrt{2} = 0.71$.

EXAMPLE 2.12.6—The following sampling experiment shows how the Central Limit Theorem performs with a population simulating what is called a U-shaped distribution. In the random digits table, score 0, 1, 2, 3 as 0; 4, 5 as 1; and 6, 7, 8, 9 as 2. In this population, the probabilities of score of 0, 1, 2 and 0.4, 0.2, and 0.4, respectively. This is a discrete distribution in which the central ordinate, 0.2, is lower than the two outside ordinates, 0.4. Draw a number of samples of size 5, using the random digits table. Record the total score for each sample. The distribution of total scores will be found fairly similar to the bell-shaped normal curve. The theoretical distribution of the total scores is as follows:

Score	0 or 10	1 or 9	2 or 8	3 or 7	4 or 6	5
Prob.	.010	.026	.077	.115	.182	.179

That is, the probability of a 0 and that of a 10 are both 0.010.

2.13—Confidence intervals for μ when σ is known. Given a random sample of size n from a population, where n is large enough so that \bar{X} can be assumed normally distributed, we are now in a position to make an interval estimate of μ . For simplicity, we assume in this section that σ is known. This is not commonly so in practice. In some situations, however, previous populations similar to the one now being investigated all have about the same standard deviation, which is known from these previous results. Further, the value of σ can sometimes be found from theoretical considerations about the nature of the population.

We first show how to find a 95% confidence interval. In section 2.1 it was pointed out that if a variate X is drawn from a normal distribution, the probability is about 0.95 that X lies between $\mu - 2\sigma$ and $\mu + 2\sigma$. More exactly, the limits corresponding to a probability 0.95 are $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$. Apply this result to \bar{X} , remembering that in repeated sampling \bar{X} has a standard deviation σ/\sqrt{n} . Thus, unless an unlucky 5% chance has come off, \bar{X} will lie between $\mu - 1.96\sigma/\sqrt{n}$ and $\mu + 1.96\sigma/\sqrt{n}$. Expressing this as a pair of inequalities, we write

$$\mu - 1.96\sigma/\sqrt{n} \leq \bar{X} \leq \mu + 1.96\sigma/\sqrt{n}$$

apart from a 5% chance. These inequalities can be rewritten so that they provide limits for μ when we know \bar{X} . The left-hand inequality is equivalent to the statement that

$$\mu \leq \bar{X} + 1.96\sigma/\sqrt{n}$$

In the same way, the right-hand inequality implies that

$$\mu \geq \bar{X} - 1.96\sigma/\sqrt{n}$$

Putting the two together, we reach the statement that unless an unlucky 5% chance occurred in drawing the sample,

$$\bar{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96\sigma/\sqrt{n}$$

This is the 95% confidence interval for μ .

Similarly, the 99% confidence interval for μ is

$$\bar{X} - 2.58\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 2.58\sigma/\sqrt{n}$$

because the probability is 0.99 that a normal deviate Z lies between the limits -2.58 and $+2.58$.

To find the confidence interval corresponding to any confidence probability P , read from the cumulative normal table (table A 3) a value Z_P , say, such that the area given in the table is $P/2$. Then the probability that a normal deviate lies between $-Z_P$ and $+Z_P$ will be P . The confidence interval is

$$\bar{X} - Z_P\sigma/\sqrt{n} \leq \mu \leq \bar{X} + Z_P\sigma/\sqrt{n}$$

One-sided confidence limits. Sometimes we want to find only an upper limit or a lower limit for μ , but not both. A company making large batches of a chemical product might have, as part of its quality control program, a regulation that each batch be tested to ensure that it does not contain more than 25 parts per million of a certain impurity, apart from a 1 in 100 chance. The test consists of drawing out n amounts of the product from the batch, and determining the concentration of impurity in each amount. If the batch is to pass the test, the 99% upper confidence limit for μ must be not more than 25 parts per million. Similarly, certain roots of tropical trees are a source of a potent insecticide whose concentration varies considerably from root to root. The buyer of a large shipment of these roots wants a guarantee that the concentration of the active ingredient in the shipment exceeds some stated value. It may be agreed between buyer and seller that the shipment is acceptable if, say, the 95% lower confidence limit for the average concentration μ exceeds the desired minimum.

To find a *one-sided* or *one-tailed* limit with confidence probability 95%, we want a normal deviate Z such that the area beyond Z in one tail is 0.05. In table A 3, the area from 0 to Z will be 0.45, and the value of Z

is 1.645. Apart from a 5% chance in drawing the sample,

$$\bar{X} \leq \mu + 1.645\sigma/\sqrt{n}$$

This gives, as the *lower* 95% confidence limit for μ ,

$$\mu \geq \bar{X} - 1.645\sigma/\sqrt{n}$$

The *upper* limit is $\bar{X} + 1.645\sigma/\sqrt{n}$. For 99% limit the value of Z is 2.326. For a one-sided limit with confidence probability P (expressed as a proportion), read table A 3 to find the Z that corresponds to probability $(P - 0.5)$.

2.14—Size of sample. The question: How large a sample must I take? is frequently asked by investigators. The question is not easy to answer. But if the purpose of the investigation is to estimate the mean of a population from the results of a sample, the methods in the preceding sections are helpful.

First, the investigator must state how accurate he would like his sample estimate to be. Does he want it to be correct to within 1 unit, 5 units, or 10 units, on the scale on which he is measuring? In trying to answer this question, he thinks of the purposes to which the estimate will be put, and tries to envisage the consequences of having errors of different amounts in the estimate. If the estimate is to be made in order to guide a specific business or financial decision, calculations may indicate the level of accuracy necessary to make the estimate useful. In scientific research it is often harder to do this, and there may be an element of arbitrariness in the answer finally given.

By one means or another, the investigator states that he would like his estimate to be correct to within some limit $\pm L$, say. Since the normal curve extends from minus infinity to plus infinity, we cannot guarantee that \bar{X} is certain to lie between the limits $\mu - L$ and $\mu + L$. We can, however, make the probability that \bar{X} lies between these limits as large as we please. In practice, this probability is usually set at 95% or 99%. For the 95% probability, we know that there is a 95% chance that \bar{X} lies between the limits $\mu - 1.96\sigma/\sqrt{n}$ and $\mu + 1.96\sigma/\sqrt{n}$. This gives the equation

$$1.96\sigma/\sqrt{n} = L$$

which is solved for n .

The equation requires a knowledge of σ , although the sample has not yet been drawn. From previous work on this or similar populations, the investigator guesses a value of σ . Since this guess is likely to be somewhat in error, we might as well replace 1.96 by 2 for simplicity. This gives the formula

$$n = 4\sigma^2/L^2$$

The formula for 99% probability is $n = 6.6\sigma^2/L^2$.

To summarize, the investigator must supply: (i) an upper limit L to the amount of error that he can tolerate in the estimate, (ii) the desired probability that the estimate will lie within this limit of error, and (iii) an advance guess at the population standard deviation σ . The formula for n is then very simple.

EXAMPLE 2.14.1—Find (i) the 80%, (ii) the 90% confidence limits for μ , given \bar{X} and σ . Ans. (i) $\bar{X} \pm 1.28\sigma/\sqrt{n}$, (ii) $\bar{X} \pm 1.64\sigma/\sqrt{n}$.

EXAMPLE 2.14.2—The heights of a random sample of 16 men from a population with $\sigma = 2.6$ in. are measured. What is the confidence probability that \bar{X} does not differ from μ by more than 1 in.? Ans. $P = 0.876$.

EXAMPLE 2.14.3—For the insecticide roots, the buyer wants assurance that the average content of the active ingredient is at least 8 lbs. per 100 lbs., apart from a 1-in-100 chance. A sample of 9 bundles of roots drawn from the batch gives, on analysis, $\bar{X} = 10.2$ lbs. active ingredient per 100 lbs. If $\sigma = 3.3$ lbs. per 100 lbs., find the lower 99% confidence limit for μ . Does the batch meet the specification? Ans. Lower limit = 7.6 lbs. per 100 lbs. No.

EXAMPLE 2.14.4—In the auditing of a firm's accounts receivable, 100 entries were checked out of a ledger containing 1,000 entries. For these 100 entries, the auditor's check showed that the stated total amount receivable exceeded the correct amount receivable by \$214. Calculate an upper 95% confidence limit for the amount by which the reported total receivable in the whole ledger exceeds the correct amount. Assume $\sigma = \$1.30$ in the population of the bookkeeping errors. Ans. \$2,354. Note: for an estimated population total, the formula for a one-sided upper limit for $N\mu$ is $N\bar{X} + NZ\sigma/\sqrt{n}$. Note also that you are given the sample total $n\bar{X} = \$214$.

EXAMPLE 2.14.5—When measurements are rounded to the nearest whole number, it can often be assumed that the error due to rounding is equally likely to lie anywhere between -0.5 and $+0.5$. That is, rounding errors follow a *uniform* distribution between the limits -0.5 and $+0.5$. From theory, this distribution has $\mu = 0$, $\sigma = 1/\sqrt{12} = 0.29$. If 100 independent, rounded measurements are added, what is the probability that the error in the total due to rounding does not exceed 5? Ans. $P = 0.916$.

EXAMPLE 2.14.6—In the part of a large city in which houses are rented, an economist wishes to estimate the average monthly rent correct to within $\pm \$20$, apart from a 1-in-20 chance. If he guesses that σ is about \$60, how many houses must he include in his sample? Ans. $n = 36$

EXAMPLE 2.14.7—Suppose that in the previous example the economist would like 99% probability that his estimate is correct to within \$20. Further, he learns that in a recent sample of 100 houses, the lowest rent was \$30 and the highest was \$260. Estimating σ from these data, find the sample size needed. Ans. $n = 36$. This estimate is, of course, very rough.

EXAMPLE 2.14.8—Show that if we wish to cut the limit of error from L to $L/2$, the sample size must be quadrupled. With the same L , if we wish 99% probability of being within the limit rather than 95% probability, what percentage increase in sample size is required? Ans. about 65% increase

2.15—"Student's" t -distribution. In most applications in which sample means are used to estimate population means, the value of σ is not known. We can, however, obtain an estimate s of σ from the sample data that give us the value of \bar{X} . If the sample is of size n , the estimate s is based on $(n - 1)$ degrees of freedom. We require a distribution that will enable us to compute confidence limits for μ , knowing s but not σ . Known

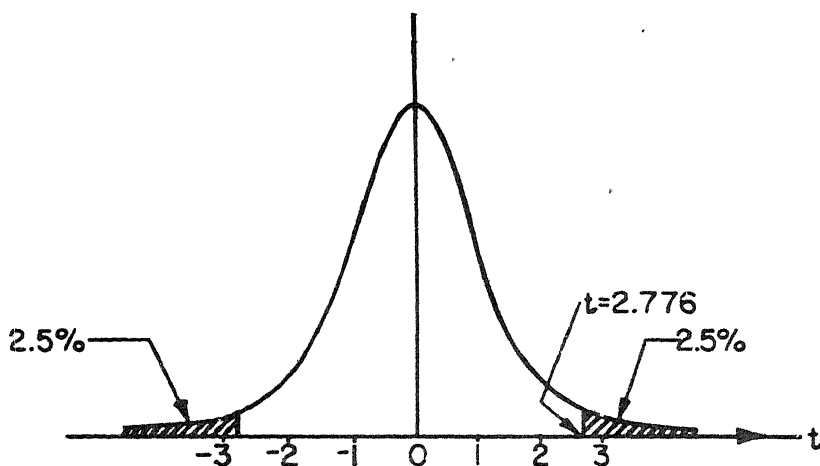


FIG. 2.15.1—Distribution of t with 4 degrees of freedom. The shaded areas comprise 5% of the total area. The distribution is more peaked in the center and has higher tails than the normal.

as “*Student’s*” t -distribution, this result was discovered by W. S. Gosset in 1908 (7) and perfected by R. A. Fisher in 1926 (8). This distribution has revolutionized the statistics of small samples. In the next chapter you will be asked to verify the distribution by the same kind of sampling process you used for chi-square; indeed, it was by such sampling that Gosset first learned about it.

The quantity t is given by the equation,

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

That is, t is the deviation of the estimated mean from that of the population, measured in terms of s/\sqrt{n} as the unit. We do not know μ though we may have some hypothesis about it. Without μ , t cannot be calculated; but its sampling distribution has been worked out.

The denominator, s/\sqrt{n} , is a useful quantity estimating σ/\sqrt{n} , the *standard error of \bar{X}* .

The distribution of t is laid out in table A 4, p. 549. In large samples it is practically normal with $\mu = 0$ and $\sigma = 1$. It is only for samples of less than 30 that the distinction becomes obvious.

Like the normal, the t -distribution is symmetrical about the mean. This allows the probability in the table to be stated as that of a larger absolute value, sign ignored. For a sample of size 5, with 4 degrees of freedom, figure 2.15.1 shows such values of t in the shaded areas; 2.5% of them are in one tail and 2.5% in the other. Effectively, the table shows the two halves of the figure superimposed, giving the sum of the shaded areas (probabilities) in both.

EXAMPLE 2.15.1—In the vitamin C sampling of table 2.8.1, $s_{\bar{x}} = 3.98/\sqrt{17} = 0.965$ mg./100 gm. Set up the hypothesis that $\mu = 17.954$ mg./100 gm. Calculate t . Ans. 2.12.

EXAMPLE 2.15.2—For the vitamin C sample, degrees of freedom $= 17 - 1 = 16$, the denominator of the fraction giving s^2 . From table A 4, find the probability of a value of t larger in absolute value than 2.12. Ans. 0.05. This means that, among random samples of $n = 17$ from normal populations, 5% of them are expected to have t -values below -2.12 or above 2.12 .

EXAMPLE 2.15.3—If samples of $n = 17$ are randomly drawn from a normal population and have t calculated for each, what is the probability that t will fall between -2.12 and $+2.12$? Ans. 0.95.

EXAMPLE 2.15.4—If random samples of $n = 17$ are drawn from a normal population, what is the probability of t greater than 2.12? Ans. 0.025.

EXAMPLE 2.15.5—What size of sample would have $t > |2|$ in 5% of all random samples from normal populations? Ans. 61. (Note the symbol for "absolute value," that is, ignoring signs.)

EXAMPLE 2.15.6—Among very large samples ($d.f. = \infty$), what value of t would be exceeded in 2.5% of them? Ans. 1.96.

2.16—Confidence limits for μ based on the t -distribution. With σ known, the 95% limits for μ were given by the relations

$$\bar{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96\sigma/\sqrt{n}$$

When σ is replaced by s , the only change needed is to replace the number 1.96 by a quantity which we call $t_{0.05}$. To find $t_{0.05}$, read table A 4 in the column headed 0.050 and find the value of t for the number of degrees of freedom in s . When the $d.f.$ are infinite, $t_{0.05} = 1.960$. With 40 $d.f.$, $t_{0.05}$ has increased to 2.021, with 20 $d.f.$ it has become 2.086, and it continues to increase steadily as the number of $d.f.$ decline.

The inequalities giving the 95% confidence limits then become

$$\bar{X} - t_{0.05} s/\sqrt{n} \leq \mu \leq \bar{X} + t_{0.05} s/\sqrt{n}$$

As illustration, recall the vitamin C determinations in table 2.8.1; $n = 17$, $\bar{X} = 20$ and $s_{\bar{x}} = 0.965$ mg./100 gm. To get the 95% confidence interval (interval estimate):

1. Enter the table with $d.f. = 17 - 1 = 16$ and in the column headed 0.05 take the entry, $t_{0.05} = 2.12$.

2. Calculate the quantity,

$$t_{0.05} s_{\bar{x}} = (2.12)(0.965) = 2.05 \text{ mg./100 gm.}$$

3. The confidence interval is from

$$20 - 2.05 = 17.95 \text{ to } 20 + 2.05 = 22.05 \text{ mg./100 gm.}$$

If you say that μ lies inside the interval from 17.95 to 22.05 mg./100 gm., you will be right unless a 1-in-20 chance has occurred in the sampling.

The point and 95% interval estimate of μ may be summarized this way: 20 ± 2.05 mg./100 gm.

62 Chapter 2: Sampling From a Normally Distributed Population

The proof of this result is similar to that given when σ is known. Although μ is unknown, the drawing of a random sample creates a value of

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

that follows Student's t -distribution with $(n - 1)$ *d.f.* Now the quantity $t_{0.05}$ in table A 4 was computed so that the probability is 0.95 that a value of t drawn at random lies between $-t_{0.05}$ and $+t_{0.05}$. Thus, there is a 95% chance that

$$-t_{0.05} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq +t_{0.05}$$

Multiply throughout by s/\sqrt{n} , and then add μ to each term in the inequalities. This gives, with 95% probability,

$$\mu - t_{0.05}s/\sqrt{n} \leq \bar{X} \leq \mu + t_{0.05}s/\sqrt{n}$$

The remainder of the proof is exactly the same as for σ known. The limits may be expressed more compactly as $\bar{X} \pm t_{0.05}s_{\bar{x}}$. For a one-sided 95% limit, use $t_{0.10}$ in place of $t_{0.05}$.

EXAMPLE 2.16.1—The yields of alfalfa from 10 plots were 0.8, 1.3, 1.5, 1.7, 1.7, 1.8, 2.0, 2.0, 2.0, and 2.2 tons per acre. Set 95% limits on the mean of the population of which this is a random sample. **Ans.** 1.41 and 1.99 tons per acre.

EXAMPLE 2.16.2—In an investigation of growth in school children in private schools, the sample mean height of 265 boys of age 13 1/2–14 1/2 years was 63.84 in. with standard deviation $s = 3.08$ in. What is the 95% confidence interval for μ ? **Ans.** 63.5 to 64.2 in.

EXAMPLE 2.16.3—In a check of a day's work for each of a sample of 16 women engaged in tedious, repetitive work, the average number of minor errors per day was 5.6, with a sample *s.d.* of 3.6. Find (i) a 90% confidence interval for the population mean number of errors, (ii) a one-sided upper 90% limit to the population number of errors. **Ans.** (i) 4.0 to 7.2 (ii) 6.8

EXAMPLE 2.16.4—We have stated that the t distribution differs clearly from the normal distribution only for samples of size less than 30. For a given value of $s_{\bar{x}}$, how much wider is (i) the 95% (ii) the 99% confidence interval when the sample size is 30 than when the sample size is very large? Are there sample sizes at which the 95% and 99% intervals become twice as wide, for the same $s_{\bar{x}}$ as with very large samples? **Ans.** (i) 4.3% wider (ii) 7.0% wider, since $s_{\bar{x}}$ has 29 *d.f.* For a sample of size 3 (2 *d.f.*) the 95% interval is twice as wide, and for a sample of size 4 the 99% interval is twice as wide. With small samples, s is not a good estimate of σ , and the confidence limits widen to allow for the chance that the sample s is far removed from the true σ .

2.17—Relative variation. Coefficient of variation. In describing the amount of variation in a population, a measure often used is the *coefficient of variation* $C = \sigma/\mu$. The sample estimate is s/\bar{X} . The standard deviation is expressed as a fraction, or sometimes as a percentage, of the mean. The utility of this measure lies partly in the fact that in many series the mean and standard deviation tend to change together. This is illustrated

by the mean stature and corresponding standard deviation of girls from 1 to 18 years of age shown graphically in figure 2.17.1. Until the twelfth year the standard deviation increases at a somewhat greater rate, relative to its mean, than does stature, causing the coefficient of variation to rise, but by the seventeenth year and thereafter C is back to where it started. Without serious discrepancy one may fix in mind the figure, $C = 3.75\%$, as the relative standard deviation of adult human stature, male as well as female. More precisely, the coefficient rises rather steadily from infancy through puberty, falls sharply during a brief period of uniformity, then takes on its permanent value near 3.75% .

A knowledge of relative variation is valuable in evaluating experiments. After the statistics of an experiment are summarized, one may judge of its success partly by looking at C . In corn variety trials, for example, although mean yield and standard deviation vary with location and season, yet the coefficient of variation is often between 5% and 15% . Values outside this interval cause the investigator to wonder if an error has been made in calculation, or if some unusual circumstances throw doubt on the validity of the experiment. Similarly, each sampler knows what values of C may be expected in his own data, and is suspicious of any great deviation. If another worker with the same type of measurement reports C values much smaller than one's own, it is worthwhile to try to discover why, since the reason may suggest ways of improving one's precision.

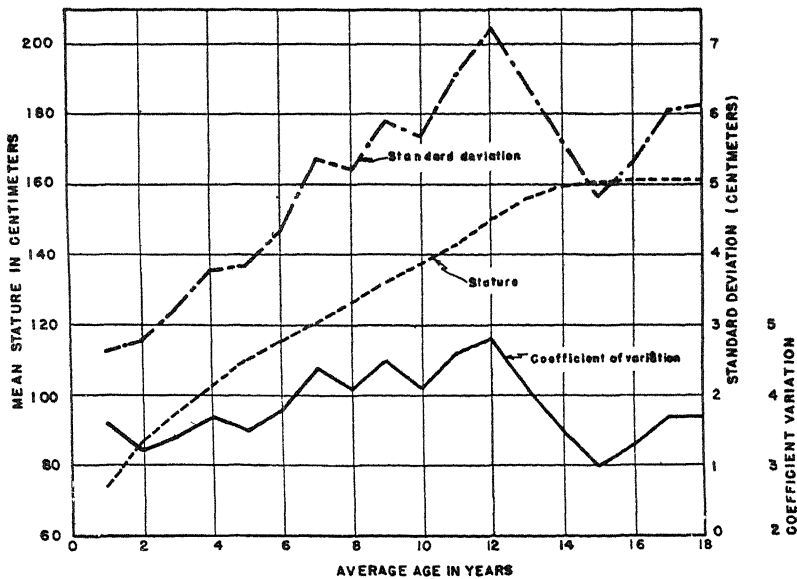


FIG 2 17 1—Graph of 3 time series, stature, standard deviation, and coefficient of variation of girls from 1 to 18 years of age. See reference (1)

Other uses of the coefficient of variation are numerous but less prevalent. Since C is the ratio of two averages having the same unit of measurement it is itself independent of the unit employed. Thus, C is the same whether inches, feet, or centimeters are used to measure height. Also, the coefficient of variation of the yield of hay is comparable to that of the yield of corn. Experimental animals have characteristic coefficients of variation, and these may be compared despite the diversity of the variables measured. Such information is often useful in guessing a value of σ for the estimation of sample size as in section 2.14.

Like many other ratios, the coefficient of variation is so convenient that some people overlook the information contained in the original data. Try to imagine how limited you would be in interpreting the stature-of-girls coefficients if they were not accompanied by \bar{X} and s . You would not know whether an increase in C is due to a rising s or a falling \bar{X} , nor whether the saw-tooth appearance of the C -curve results from irregularities in one or both of the others, unless indeed you could supply the facts from your own fund of knowledge. The coefficient is informative and useful in the presence of \bar{X} and s , but abstracted from them it may be misleading.

EXAMPLE 2 17 1—In experiments involving chlorophyll determinations in pineapple plants (10), the question was raised as to the method that would give the most consistent results. Three bases of measurement were tried, each involving 12-leaf samples, with the statistics reported below. From the coefficients of variation, it was decided that the methods were equally reliable, and the most convenient one could be chosen with no sacrifice of precision.

STATISTICS OF CHLOROPHYLL DETERMINATIONS OF 12-LEAF SAMPLES FROM PINEAPPLE PLANTS, USING THREE BASES OF MEASUREMENT

Statistic	100-gram Wet Basis	100-gram Dry Basis	100-sq. cm. Basis
Sample Mean (milligrams)	61.4	337	13.71
Sample Standard Deviation (milligrams)	5.22	31.2	1.20
Coefficient of Variation (per cent)	8.5	9.3	8.8

EXAMPLE 2 17 2—In a certain laboratory there is a colony of rats in which the coefficient of variation of the weights of males between 56 and 84 days of age is close to 13%. Estimate the sample standard deviation of the weights of a lot of these rats whose sample mean weight is 200 grams. Ans. 26 grams

EXAMPLE 2 17 3—If C is the coefficient of variation in a population, show that the coefficient of variation of the mean of a random sample of size n is C/\sqrt{n} in repeated sampling. Does the same result hold for the sample total? Ans. Yes

EXAMPLE 2 17 4—If the coefficient of variation of the gain in weight of a certain animal over a month is 10%, what would you expect the coefficient of variation of the gain over a four-month period to be? Ans. The answer is complicated, and cannot be given fully at this stage. If σ and μ were the same during each of the four months, and if the gains were independent from month to month, the answer would be $C/\sqrt{4} = C/2$, by the result in the preceding example. But animals sometimes grow by spurts, so that the gains in

successive periods may not be independent, and our formula for the standard deviation of a sample does not apply in this case. The answer is likely to lie between C and $C/2$. The point will be clarified when we study correlation.

REFERENCES

- 1 COUNCIL ON FOODS. *JAMA*, 110:651 (1938).
- 2 E. S. PEARSON. *Biometrika*, 24:416 (1932).
- 3 L. H. C. TIPPETT. *Biometrika*, 17:386 (1925).
- 4 A. R. CUSHNY and A. R. PEEBLES. *Amer. J. Physiol.*, 32:501 (1905).
- 5 A. M. MOOD and F. A. GRAYBILL. *Introduction to the Theory of Statistics*, 2nd ed. McGraw-Hill, New York (1963).
6. *Statistical Abstract of the United States*. U.S. GPO, Washington, D.C (1959).
- 7 "Student" *Biometrika*, 6:1 (1908).
- 8 R. A. FISHER. *Metron*, 5:90 (1926).
- 9 P. J. TALLEY. *Plant Physiol*, 9:737 (1934)
- 10 R. K. TAM and O. C. MAGISTAD. *Plant Physiol*, 10:161 (1935)

Experimental sampling from a normal population

3.1—Introduction. In chapter 1 the facts about confidence intervals for a proportion were verified through experimental sampling. This same device illustrated the theoretical distribution of chi-square that forms the basis of the test of a null hypothesis about the population proportion. In chapter 2 the results of two experimental samplings were presented to show that the distribution of means of random samples tends to approximate the normal distribution with standard deviation σ/\sqrt{n} , as predicted by the Central Limit Theorem.

In this chapter we present further experimental samplings from a population simulating the normal, with instructions so that the reader can perform his own samplings. The purposes are as follows:

- (1) To provide additional verification of the result that the sample means are normally distributed with S.D. = σ/\sqrt{n} .
- (2) To investigate the sampling distribution of s^2 , regarded as an estimate of σ^2 , and of s , regarded as an estimate of σ . Thus far we have not been much concerned with the question: How good an estimate of σ^2 is s^2 ? The frequency distribution of s^2 in normal samples has, however, been worked out and tabulated. Apart from a multiplier, it is an extended form of the chi-square distribution which we met in chapter 1.
- (3) To illustrate the sampling distribution of t with 9 degrees of freedom, by comparing the values of t found in the experimental sampling with the theoretical distribution.
- (4) To verify confidence interval statements based on the t -distribution.

The population that we have devised to simulate a normal population departs from it in two respects: it is limited in size and range instead of being infinite, and has a *discontinuous* variate instead of the continuous one implied in the theory. The effects of these departures will scarcely be noticed, because they are small in comparison with sampling variation.

3.2—A finite population simulating the normal. In table 3.2.1 are the weight gains of a hundred swine, slightly modified from experimental data so as to form a distribution which is approximately normal with

TABLE 3.2.1
 ARRAY OF GAINS IN WEIGHT (POUNDS) OF 100 SWINE DURING A PERIOD OF 20 DAYS
 The gains approximate a normal distribution with
 $\mu = 30$ pounds and $\sigma = 10$ pounds

Item Number	Gain	Item Number	Gain	Item Number	Gain	Item Number	Gain
00	3	25	24	50	30	75	37
01	7	26	24	51	30	76	37
02	11	27	24	52	30	77	38
03	12	28	25	53	30	78	38
04	13	29	25	54	30	79	39
05	14	30	25	55	31	80	39
06	15	31	26	56	31	81	39
07	16	32	26	57	31	82	40
08	17	33	26	58	31	83	40
09	17	34	26	59	32	84	41
10	18	35	27	60	32	85	41
11	18	36	27	61	33	86	41
12	18	37	27	62	33	87	42
13	19	38	28	63	33	88	42
14	19	39	28	64	33	89	42
15	19	40	28	65	33	90	43
16	20	41	29	66	34	91	43
17	20	42	29	67	34	92	44
18	21	43	29	68	34	93	45
19	21	44	29	69	35	94	46
20	21	45	30	70	35	95	47
21	22	46	30	71	35	96	48
22	22	47	30	72	36	97	49
23	23	48	30	73	36	98	53
24	23	49	30	74	36	99	57

$\mu = 30$ pounds and $\sigma = 10$ pounds. The items are numbered from 00 to 99 in order that they may be identified easily with corresponding numbers taken from the table of random digits. The salient features of this kind of distribution may be discerned in figure 3.2.1. The gains, clustering at the midpoint of the array, thin out symmetrically, slowly at first, then more and more rapidly: two-thirds of the gains lie in the interval 30 ± 10 pounds, that is, in an interval of two standard deviations centered on the mean. In a real population, indefinitely great in number of individuals, greater extremes doubtless would exist, but that need cause us little concern.

The relation of the histogram to the array is clear. After the class bounds are decided upon, it is necessary merely to count the dots lying between the vertical lines, then make the height of the rectangle proportional to their number. The central value, or *class mark*, of each interval is indicated on the horizontal scale of gains.

In table 3.2.2 is the frequency distribution which is graphically represented in figure 3.2.1. Only the class marks are entered in the first row. The class intervals are from 2.5 to 7.5, etc.

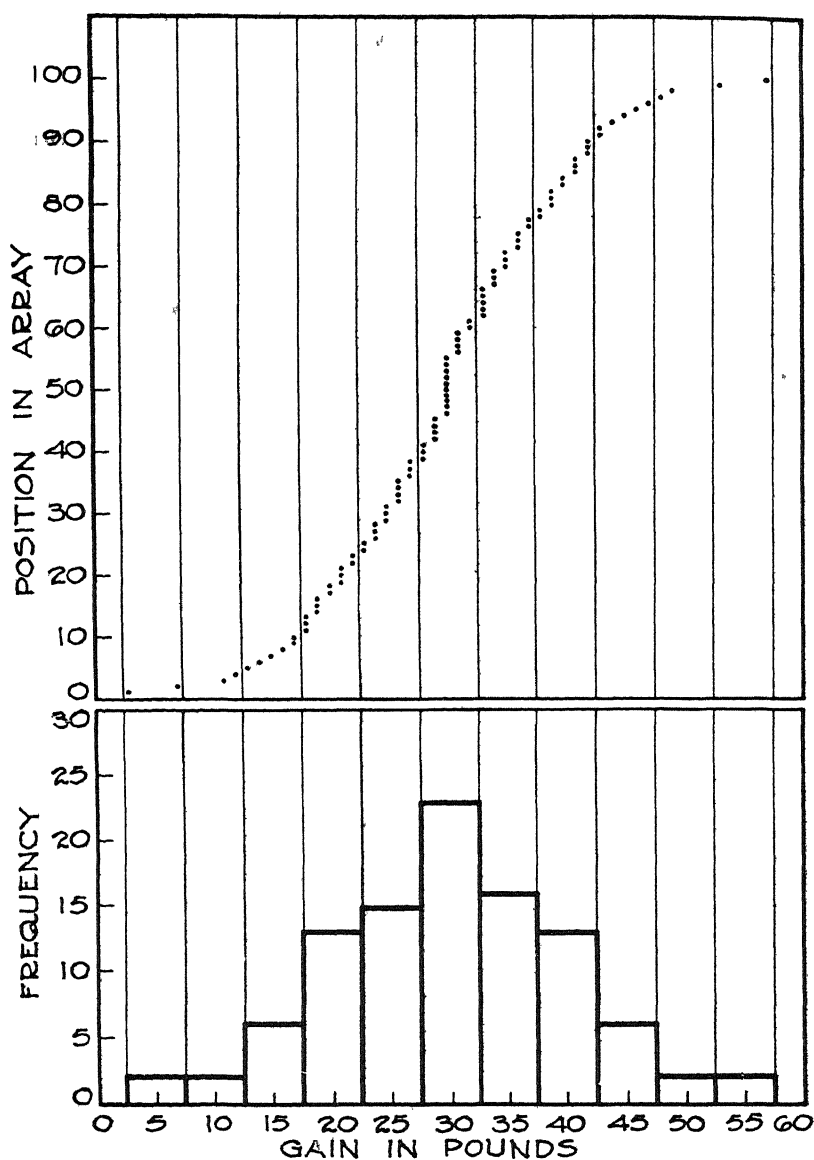


FIG 3.2.1—Upper part Graphical representation of array of 100 normally distribute gains in weight. Lower part Histogram of same gains. The altitude of a rectangle in the histogram is proportional to the number of dots in the array which lie between the vertical sides.

TABLE 3.2.2
FREQUENCY DISTRIBUTION OF GAINS IN WEIGHT OF 100 SWINE
(A finite population approximating the normal)

Class mark (pounds)	5	10	15	20	25	30	35	40	45	50	55
Frequency	2	2	6	13	15	23	16	13	6	2	2

3.3—Random samples from a normal distribution. An easy way to draw random samples from the table of pig gains is to take numbers consecutively from the table of random numbers, table A 1, then match them with the gains by means of the integers, 00 to 99, in table 3.2.1. To avoid duplicating the samples of others in class work, start at some randomly selected point in the table of random numbers instead of at the beginning, then proceed upward, downward, or crosswise. Suppose you have hit upon the digit, 8, in row 71, column 29. This, with the following digit, 3, specifies pig number 83 in table 3.2.1, a pig whose gain is 40 pounds. Hence, 40 pounds is the first number of the sample. Moving upward among the random numbers you read the integers 09, 75, 90, etc., and record the corresponding gains from the table, 17, 37, and 43 pounds. Continuing, you get as many gains and as many samples as you wish.

Samples of 10 are suggested. For our present purposes all the samples must be of the same size because the distributions of their statistics

TABLE 3.3.1
FOUR SAMPLES OF 10 ITEMS DRAWN AT RANDOM FROM THE PIG GAINS OF TABLE 3.2.1,
EACH FOLLOWED BY STATISTICS TO BE EXPLAINED IN SECTIONS 3.4–3.8

Item Number and Formulas	Sample Number			
	1	2	3	4
1	33	32	39	17
2	53	31	34	22
3	34	11	33	20
4	29	30	33	19
5	39	19	33	3
6	57	24	39	21
7	12	53	36	3
8	24	44	32	25
9	39	19	32	40
10	36	30	30	21
\bar{X}	35.6	29.3	34.1	19.1
s^2	169.8	151.6	9.0	112.3
s	13.0	12.3	3.0	10.6
$s_{\bar{X}}$	4.11	3.89	0.95	3.35
t	1.36	-0.18	4.32	-3.25
$t_{0.05} s_{\bar{X}}$	9.3	8.8	2.2	7.6
$\bar{X} \pm t_{0.05} s_{\bar{X}}$	26.3–44.9	20.5–38.1	31.9–36.3	11.5–26.7

change with n . It is well to record the items in columns, leaving a half dozen lines below each for subsequent computations. For your guidance, four samples are listed in table 3.3.1. The computations below them will be explained as we go along. Draw as many of the samples as you think you can process within the time at your command. If several are working together the results of each can be made available to all. Keep the records carefully because you will need them again and again.

Each pig gain may be drawn as often as its number appears in the table of random digits—it is not withdrawn from circulation after being taken once. Thus, the sampling is always from the same population, and the probability of drawing any particular item is constant throughout the process.

EXAMPLE 3.3.1—Determine the range in each of your samples of $n = 10$. The mean of the ranges estimates $\sigma/0.325$ (table 2 4 1); that is, $10/0.325 = 30.8$. How close is your estimate?

3.4.—The distribution of sample means. First add the items in each sample, then put down the sample mean, \bar{X} (division is by 10). While every mean is an estimator of $\mu = 30$ pounds, there is yet great variation among them. Make an array of the means of all your samples. If there are enough of them, group them into a frequency distribution like table 3.4.1.

Our laboratory means ranged from 19 to 39 pounds, perhaps to the novice a disconcerting variability. To assess the meaning of this, try to imagine doing an experiment resulting in one of these more divergent mean gains instead of the population value, 30 pounds. Having no information about the population except that furnished by the sample, you would be considerably misled. There is no way to avoid this hazard. One of the objects of the experimental samplings is to acquaint you with the risks involved in all conclusions based on small portions of the aggregate. The investigator seldom knows the parameters of the sampled population; he knows only the sample estimates. He learns to view his experimental data in the light of his experience of sampling error. His judgments must involve not only the facts of his sample but all the related information which he and others have accumulated.

The more optimistic draw satisfaction from the large number of means near the center of the distribution. If this were not characteristic, sampling would not be so useful and popular. The improbability of getting poor estimates produces a sense of security in making inferences.

Fitting the normal distribution. In constructing table 3.4.1, one-pound class intervals were used. Since all the means come out exactly to one decimal place, the class limits were taken as 19.5–20.4, 20.5–21.4, and so on.

From theory, the distribution of sample means should be very close to normal, with mean $\mu = 30$ pounds and standard deviation $\sigma_{\bar{X}} = 10/\sqrt{10} = 3.162$ pounds. The theoretical frequencies appear in the right-hand

TABLE 3.4.1
FREQUENCY DISTRIBUTION OF 511 MEANS OF SAMPLES OF 10 DRAWN FROM
THE PIG GAINS IN TABLE 3.2.1

Class Limits (Pounds)	Observed Frequency	Theoretical Frequency
Less than 19.5	1	0.20
19.5-20.4	1	0.46
20.5-21.4	0	1.12
21.5-22.4	7	2.56
22.5-23.4	5	5.47
23.5-24.4	10	10.48
24.5-25.4	19	18.09
25.5-26.4	30	28.46
26.5-27.4	41	40.52
27.5-28.4	48	52.12
28.5-29.4	66	60.76
29.5-30.4	72	64.18
30.5-31.4	56	61.32
31.5-32.4	46	53.25
32.5-33.4	45	41.65
33.5-34.4	22	29.59
34.5-35.4	24	19.11
35.5-36.4	12	11.09
36.5-37.4	5	5.88
37.5-38.4	0	2.76
Over 38.5	1	1.94
Total	511	511.01

column of table 3.4.1. To indicate how these are computed, let us check the frequency 28.46 for the class whose limits are 25.5-26.4. First we must take note of the fact that our computed means are *discrete*, since they change by intervals of 0.1, whereas the normal distribution is *continuous*. No computed mean in our samples can have a value of, say, 25.469, although the normal distribution allows such values. This discrepancy is handled by regarding any discrete mean as a grouping of all continuous values to which it is nearest. Thus, the observed mean of 25.5 represents all continuous values lying between 25.45 and 25.55. Similarly, the observed mean 26.4 represents the continuous values between 26.35 and 26.45. Hence for the class whose discrete limits are 25.5 and 26.4, we take the true class limits as 25.45 and 26.45. When fitting a continuous theoretical distribution to an observed frequency distribution, the true class limits must always be found in this way.

In order to use the normal table, we express the true limits in standard measure. For $\bar{X} = 25.45$, $\mu = 30$, $\sigma_{\bar{X}} = 3.162$, we have

$$Z_1 = (\bar{X} - \mu)/\sigma_{\bar{X}} = (25.45 - 30)/3.162 = -1.439$$

For $\bar{X} = 26.45$, we find $Z_2 = -1.123$. From table A 3 (p. 548) we read the area of the normal curve between -1.123 and -1.439 . By symmetry,

this is also the area between 1.123 and 1.439. Linear interpolation in the table is required. The area from 0 to 1.43 is 0.4236 and from 0 to 1.44 is 0.4251. Hence, by linear interpolation, the area from 0 to 1.439 is

$$(0.9)(0.4251) + (0.1)(0.4236) = 0.4250.$$

Similarly, the area from 0 to 1.123 is 0.3693 so that the required area is 0.0557. Finally, since there are 511 means in the frequency distribution, the theoretical frequency in this class is $(511)(0.0557) = 28.46$.

To summarize, the steps in fitting a normal distribution are: (i) Find the true class limits. (ii) Express each limit in standard measure, getting a series of values Z_1, Z_2, Z_3, \dots (iii) From table A 3, read the areas from 0 to Z_1 , 0 to Z_2 , 0 to Z_3, \dots (iv) The theoretical probabilities in the classes are the areas from $-\infty$ to Z_1 , from Z_1 to Z_2 , from Z_2 to Z_3 , and so on, ending with the area from Z_k to $+\infty$, where Z_k is the lower limit of the highest class. The area from $-\infty$ to Z_1 is $0.5 -$ (area from 0 to Z_1), and the area from Z_k to $+\infty$ is $0.5 -$ (area from 0 to Z_k). The intermediate areas are all found by subtraction as in the numerical illustration. The only exception is the area that straddles the mean, say from Z_u to Z_{u+1} . Here, Z_u will be negative and Z_{u+1} positive. In this case we add the area from 0 to Z_u and that from 0 to Z_{u+1} . (v) Finally, multiply each area by the total observed frequency.

If you have used the same class limits as in table 3.4.1 but have drawn a different number of samples, say 200, multiply the theoretical frequencies in table 3.4.1 by 200/511 to obtain your comparable theoretical frequencies. If you used two-pound classes, as is advisable with a smaller number of samples, add the theoretical frequencies in table 3.4.1 in appropriate pairs and multiply by the relative sample sizes.

It is clear from table 3.4.1 that the observed frequencies are a good fit to the theoretical frequencies.

3.5—Sampling distributions of s^2 and s . For each sample, calculate s^2 by the shortcut formula,

$$s^2 = \{\Sigma X^2 - (\Sigma X)^2/10\}/9$$

Four values of s^2 are shown in table 3.3.1. Three of them overestimate $\sigma^2 = 100$, while the fourth is notably small. Examine any of your samples with unusual s^2 to learn what peculiarities of the sample are responsible. The freakish sample 3 in the table has a range of only $39-30 = 9$ pounds, with not a single member less than μ . This sample gave the smallest s^2 that appeared in our set of 511 values.

The distribution of s^2 in our 511 samples is displayed in table 3.5.1. Notice its *skewness*, with bunching below the mean and a long tail above—resembling the chi-square distribution of chapter 1, though less extreme. Despite this, the mean of the values of s^2 is 101.5, closely approximating the population variance, 100, and verifying the fact that s^2 is an unbiased estimator of σ^2 .

TABLE 3.5.1
OBSERVED AND THEORETICAL DISTRIBUTIONS OF 511 MEAN SQUARES, s^2 , OF NORMAL
SAMPLES WITH $n = 10$

Class Mark	20	40	60	80	100	120	140	160	180	200	220	240	260	280	300	320	340
Frequency Obs	12	47	92	93	72	73	42	29	26	11	8	2	1	0	1	1	1
Theor	12.8	50.8	84.8	94.7	84.5	65.2	45.9	29.6	18.4	10.8	6.1	3.4			3.8*		

* Combined frequency in 5 classes

Our distribution of s , shown in table 3.5.2, has a slight skewness (not as large as that of s^2) as well as a small bias, with mean 9.8 pounds, slightly less than $\sigma = 10$ pounds. Even in samples as small as 10 the bias is unimportant in a single estimate s .

TABLE 3.5.2
FREQUENCY DISTRIBUTION OF 511 SAMPLE STANDARD DEVIATIONS CORRESPONDING TO
* THE MEAN SQUARES OF TABLE 3.5.1

Class mark	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Frequency	1	2	9	18	58	77	80	71	79	44	41	17	8	3	1	2

The theoretical distribution of s^2 . We have already mentioned that the distribution of s^2 in normal samples is closely related to the chi-square distribution. First, we give a general definition of the chi-square distribution. If Z_1, Z_2, \dots, Z_f are independently drawn random normal deviates, the quantity

$$\chi^2 = Z_1^2 + Z_2^2 + \dots + Z_f^2$$

follows the chi-square distribution with f degrees of freedom. Thus, chi-square with f degrees of freedom is defined as the distribution followed by the sum of squares of f independent normal deviates. The form of this distribution was worked out mathematically. It could, alternatively, be examined by experimental sampling. By expressing the 100 gains in table 3.2.1 in standard measure, we would have a set of normal deviates from which we could draw samples of size f , computing χ^2 as defined above for each sample. For more accurate work, there are tables of random normal deviates (1)(2), that provide a basis for such samplings. Table A 5 (p. 550) presents the percentage points of the χ^2 distribution. It will be much used at various points in this book.

A second result from theory is that if s^2 is a mean square with f degrees of freedom, computed from a normal population that has variance σ^2 , then the quantity fs^2/σ^2 follows the chi-square distribution with f degrees of freedom. This is an exact mathematical result. Since our sample variances have $(n - 1) d.f.$, the relation is

$$\chi^2 = (n - 1)s^2/\sigma^2$$

We cannot present a proof of this result, but a little algebra makes the relation between s^2 and χ^2 clearer. Remember that $(n-1)s^2$ is the sum of squares of deviations, $\Sigma(X - \bar{X})^2$. Introduce μ as a working mean. From the identity for working means (section 2.10) we have

$$\frac{(n-1)s^2}{\sigma^2} = \frac{(X_1 - \mu)^2}{\sigma^2} + \frac{(X_2 - \mu)^2}{\sigma^2} + \dots + \frac{(X_n - \mu)^2}{\sigma^2} - \frac{n(\bar{X} - \mu)^2}{\sigma^2}$$

Now, the quantities $(X_1 - \mu)/\sigma, (X_2 - \mu)/\sigma, \dots, (X_n - \mu)/\sigma$, are all in standard measure: in other words, they are random normal deviates. And the quantity $\sqrt{n}(\bar{X} - \mu)/\sigma$ is another normal deviate, since the standard deviation of \bar{X} is σ/\sqrt{n} . Hence we may write

$$\frac{(n-1)s^2}{\sigma^2} = Z_1^2 + Z_2^2 + \dots + Z_n^2 - Z_{n+1}^2$$

Thus, $(n-1)s^2/\sigma^2$ is the sum of squares of n normal deviates, *minus* the square of one normal deviate, whereas χ^2 , with $(n-1)$ d.f., is the sum of the squares of $(n-1)$ normal deviates. It is not difficult to show mathematically that these two distributions are the same in this case.

The theoretical frequencies for our 511 values of s^2 appear in the last line of table 3.5.1. Again, the agreement with the observed frequencies is good. For fitting this distribution, table A 5 is not very convenient. We used the table in reference (3), which gives, for specified values of χ^2 , the probability of exceeding the specified value.

From the definition of the chi-square distribution, we see that chi-square with 1 degree of freedom is the distribution followed by the square of a *single* normal deviate. Later (chapter 8) we shall show that the chi-square test criterion which we encountered in chapter 1 when testing a proportion is approximately distributed as the square of a normal deviate.

Like the normal distribution, the theoretical distribution of chi-square is continuous. Unlike the normal, χ^2 , being a sum of squares, cannot take negative values, so that the distribution extends from 0 to ∞ , whereas the normal, of course, extends from $-\infty$ to $+\infty$. An important result from theory is that the mean value of χ^2 with f degrees of freedom is exactly f . Since $s^2 = \chi^2 \sigma^2 / f$, a consequence of this result is that the mean value of s^2 , in its theoretical distribution, is exactly σ^2 . This verifies the result mentioned in chapter 2 when we stated that s^2 is an unbiased estimator of σ^2 . The property that s^2 is unbiased does not require normality, but only that the sample be a random sample.

3.6—Interval estimates of σ^2 . With continuous populations, our attention thus far has centered on the problem of estimating the population *mean* from a sample. In studying the precision of measuring instruments and in studying variability in populations, we face the problem of estimating the population variance σ^2 from a sample. If the population is

normal, the χ^2 table can be used to compute a confidence interval for σ^2 from a sample value s^2 .

The entries in the chi-square table (p. 550) are the values of χ^2 that are exceeded with the probabilities stated at the heads of the columns. For a 95% confidence interval, the relevant quantities are $\chi^2_{0.975}$, the value of the chi-square exceeded with probability 0.975, and $\chi^2_{0.025}$, the value of chi-square exceeded with probability 0.025. Hence, the probability that a value of χ^2 drawn at random lies between these two limits is $0.975 - 0.025 = 0.95$. Since $\chi^2 = fs^2/\sigma^2$, the probability is 95% that when our sample was drawn,

$$\chi^2_{0.975} \leq \frac{fs^2}{\sigma^2} \leq \chi^2_{0.025}$$

Multiplying through by σ^2 , we have

$$\sigma^2 \chi^2_{0.975} \leq fs^2 \leq \sigma^2 \chi^2_{0.025}$$

The reader may verify that these inequalities are equivalent to the following,

$$\frac{fs^2}{\chi^2_{0.025}} \leq \sigma^2 \leq \frac{fs^2}{\chi^2_{0.975}}$$

This is the general formula for 95% confidence limits. With s^2 computed from a sample of size n , we have $f = (n - 1)$, and fs^2 is the sum of squares of deviations, Σx^2 . The simplest form for computing is, therefore,

$$\frac{\Sigma x^2}{\chi^2_{0.025}} \leq \sigma^2 \leq \frac{\Sigma x^2}{\chi^2_{0.975}}$$

As an illustration we shall set confidence limits on σ^2 for the population of vitamin C concentrations sampled in section 2.4. For these data, $\Sigma x^2 = 254$, $d.f. = 16$, $s^2 = 15.88$. From table A 5, $\chi^2_{0.975} = 6.91$ and $\chi^2_{0.025} = 28.8$. Substituting,

$$\frac{254}{28.8} \leq \sigma^2 \leq \frac{254}{6.91},$$

that is,

$$8.82 \leq \sigma^2 \leq 36.76,$$

gives the confidence interval for σ^2 . Unless a 1-in-20 chance has occurred in the sampling, σ^2 lies between 8.82 and 36.76. To obtain confidence limits for σ , take the square roots of these limits. The limits for σ are 2.97 and 6.06 mg./100 gm. Note that $s = 3.98$ is not in the middle of the interval, since the distribution of s is skew.

Large samples are necessary if σ is to be estimated accurately. For illustration, assume that by an accurate estimate we mean one that is known, with confidence probability 95%, to be correct to within $\pm 10\%$. If our estimate s is 100, the confidence limits for σ should be 90 and 110. Consider a sample of size 101, giving 100 *d.f.* in s^2 . From the last line of table A 5, with $s^2 = 10,000$, the 95% limits for σ^2 are 7,720 and 13,470, so that those for σ are 87.9 and 116. Thus, even a sample of 101 does not produce limits that are within 10% of the estimated value. For a sample of size 30, with $s = 100$, the limits are 80 and 134. The estimate could be in error by more than 20%.

The frequency distribution of s^2 is sensitive to non-normality in the original population, and can be badly distributed by any gross errors that occur in the sample. This effect of non-normality is discussed further in section 3.15.

3.7—Test of a null hypothesis value of σ^2 . Situations in which it is necessary to test whether a sample value of s^2 is consistent with a postulated population value of σ^2 are not too frequent in practice. This problem does arise, however, in some applications in which σ^2 has been obtained from a very large sample and may be assumed known. In others, in genetics for example, a value of σ^2 may be predicted from a theory that is to be tested. The following examples indicate how the test is made.

Let the null hypothesis value of σ^2 be σ_0^2 . Usually, the tests wanted are one-tailed tests. When the alternative is $\sigma^2 > \sigma_0^2$, compute

$$\chi^2 = \frac{fs^2}{\sigma_0^2} = \frac{\Sigma x^2}{\sigma_0^2}$$

This value is significant, at the 5% level, if it exceeds $\chi^2_{0.050}$ with f degrees of freedom. Suppose that an investigator has used for years a stock of inbred rats whose weights have $\sigma_0 = 26$ grams. He considers switching to a cheaper source of supply of rats, except that he suspects that the new rats will show greater variability. An experiment on 20 new rats gave $\Sigma x^2 = 23,000$, $s = 35$ grams, in line with his suspicions. As a check he tests the null hypothesis: $\sigma = 26$ grams, against the alternative: $\sigma > 26$ grams.

$$\chi^2 = \frac{23,000}{(26)^2} = 34.02, \quad d.f. = 19$$

In table A 5, $\chi^2_{0.050}$ is 30.14, so that the null hypothesis is rejected.

To test $H_1: \sigma^2 < \sigma_0^2$, reject at the 5% level if $\chi^2 < \chi^2_{0.950}$. To illustrate, a standard method of performing an intricate chemical analysis gives $\sigma_0 = 4.9$ parts per 1,000 for the content of some chemical constituent. A refinement on the analysis, which may improve the precision and cannot make it worse, gave $s = 4.1$, based on 49 *d.f.* We have

$\chi^2 = (49)(4.1)^2/(4.9)^2 = 34.3$. Table A 5 gives $\chi^2 = 34.76$ for $f = 50$ and 26.51 for $f = 40$. Interpolating linearly, we find $\chi^2_{0.950} = 33.9$ for $f = 49$. Formally, the null hypothesis would not be rejected, though the significance probability is very close to 5%.

If H_A is the two-sided alternative $\sigma^2 \neq \sigma_0^2$, the region of rejection is $\chi^2 < \chi^2_{0.975}$ and $\chi^2 > \chi^2_{0.025}$.

EXAMPLE 3.7.1—For the fitted normal distribution in table 3.4.1, verify the theoretical frequencies (i) 1.94 for the class "Over 38.5" and (ii) 64.18 for the class "29.5–30.4."

EXAMPLE 3.7.2—If half the standard deviations in table 3.5.2 were expected to be less than $\sigma = 10$ pounds, as would be true if s were symmetrically distributed about σ , calculate $\chi^2 = 4.89$, with 1 *d.f.* for the sample. The fact that χ^2 is significant is evidence against a symmetrical distribution in the population.

EXAMPLE 3.7.3—In a sample of 61 patients, the amount of an anesthetic required to produce anesthesia suitable for surgery was found to have a standard deviation (from patient to patient) of $s = 10.2$ mg. Compute 90% confidence limits for σ . Ans. 8.9 and 12.0 mg. Use $\chi^2_{0.950}$ and $\chi^2_{0.050}$.

EXAMPLE 3.7.4—With routine equipment like light bulbs, which wear out after a time, the standard deviation of the length of life is an important factor in determining whether it is cheaper to replace all the pieces at fixed intervals or to replace each piece individually when it breaks down. For a certain gadget, an industrial statistician has calculated that it will pay to replace at fixed intervals if $\sigma < 6$ days. A sample of 71 pieces gives $s = 4.2$ days. Examine this question (i) by finding the upper 95% limit for σ from s , (ii) by testing the null hypothesis $\sigma = \sigma_0 = 6$ days against the alternative $\sigma < 6$ days. Ans. (i) The upper 95% limit is 5.0 (ii) H_0 is rejected at the 5% level. Notice that the two procedures are equivalent; if the upper confidence limit had been 6.0 days, the chi-square value would be at the 5% significance level.

EXAMPLE 3.7.5—For *d.f.* greater than 100, which are not shown in table A 5, an approximation due to R. A. Fisher is that $\sqrt{2\chi^2}$ is normally distributed with mean $\sqrt{2f-1}$ and standard deviation 1. Check this approximation by finding the value that it gives for $\chi^2_{0.025}$ when $f = 100$, the correct value being 129.56. Ans. 129.1.

3.8—The distribution of t . Returning to our experimental samples, we are ready to examine the t -distribution for 9 degrees of freedom. Since \bar{X} and $s_{\bar{X}}$ have already been calculated for each of your samples of 10, the sample value of t may now be got by putting $\mu = 30$, the formula being

$$t = (\bar{X} - 30)/s_{\bar{X}}$$

Here, t will be positive or negative according as \bar{X} is greater or less than 30 pounds. In the present sampling the two signs are equally likely, so you may expect about half of each. On account of this symmetry the mean of all your t should be near zero.

The four samples in table 3.3.1 were selected to illustrate the manner in which large, small, and intermediate values of t arise in sampling. A small deviation, $(\bar{X} - \mu)$, or a large sample standard error tend to make t small. Some striking combinations are put in the table, and you can doubtless find others among your samples.

TABLE 3.8.1
SAMPLE AND THEORETICAL DISTRIBUTIONS OF t . SAMPLES OF 10
DEGREES OF FREEDOM, 9

Interval of t		Sample		Theoretical		
From	To	Frequency	Percentage Frequency	Percentage Frequency	Cumulative	
					One Tail	Both Tails
.....	-3.250	3	0.6	0.5	100.0	
-3.250	-2.821	4	0.8	0.5	99.5	
-2.821	-2.262	5	1.0	1.5	99.0	
-2.262	-1.833	16	3.1	2.5	97.5	
-1.833	-1.383	31	6.1	5.0	95.0	
-1.383	-1.100	25	4.9	5.0	90.0	
-1.100	-0.703	52	10.2	10.0	85.0	
-0.703	0.0	132	25.8	25.0	75.0	
0.0	0.703	126	24.6	25.0	50.0	100.0
0.703	1.100	41	8.0	10.0	25.0	50.0
1.100	1.383	32	6.3	5.0	15.0	30.0
1.383	1.833	18	3.5	5.0	10.0	20.0
1.833	2.262	13	2.5	2.5	5.0	10.0
2.262	2.821	8	1.6	1.5	2.5	5.0
2.821	3.250	2	0.4	0.5	1.0	2.0
3.250		3	0.6	0.5	0.5	1.0
		511	100.0	100.0		

The distribution of the laboratory sample of t is displayed in table 3.8.1. The class intervals in the present table are unequal, adjusted so as to bring into prominence certain useful probabilities in the tails of the distribution. The theoretical percentage frequencies are recorded for comparison with those of the sample. The agreement is remarkably good. In the last two columns are the cumulative percentage frequencies which make the table convenient for confidence statements and tests of hypotheses. Examination of the table reveals that 2.5% of all t -values in samples of 10 theoretically fall beyond 2.262, while another 2.5% of values are smaller than -2.262. Combining these two tails of the distribution, as shown in the last column, 5% of all t in samples of 10 lie further from the center than $|2.262|$, which is therefore the 5% level of t . Make a distribution of your own sample t to be compared with the theoretical distributions in the table.

Our t -table, table A 4, is a two-tailed table because most applications of the t -distribution call for two-sided confidence limits and two-tailed tests of significance. If you need a table that gives the probability for specified values of t instead of t for specified probabilities, see (4).

3.9—The interval estimate of μ ; the confidence interval. The theory of the confidence interval may now be verified from your sampling. Each

sample specifies an interval, $\bar{X} \pm t_{0.05} s_{\bar{X}}$, said to cover μ . In each of your samples, substitute the estimators, \bar{X} and $s_{\bar{X}}$, together with $t_{0.05} = 2.262$, the 0.05 level for 9 *d.f.* Finally, if you say, for any particular sample, that the interval includes μ you will be either right or wrong; which it is may be determined readily because you know that $\mu = 30$ pounds. The theory will be verified if about 95% of your statements are right and about 5% wrong.

Table 3.3.1 (p. 69) gives the steps in computing confidence limits for four samples. The intervals given by these four samples are, respectively,

26.3 to 44.9
20.5 to 38.1
31.9 to 36.3
11.5 to 26.7

Sample 1 warrants the statement that μ lies between 26.3 and 44.9 pounds, and we know that this interval does contain μ , as does likewise the interval from sample 2. On the contrary, samples 3 and 4 illustrate cases leading to false statements, one because of an unusually divergent sample mean, the other because of a small sample standard deviation. Sample 3 is particularly misleading: not only does it miss the mark, but the narrow confidence interval suggests that we have an unusually accurate estimate. Of the 511 laboratory samples, 486 resulted in correct statements about μ ; that is, 95.1% of the statements were true. The percentage of false statements, 4.9%, closely approximated the theoretical 5%. Always bear in mind the condition involved in every confidence statement at the 5% level—it is right unless a 1-in-20 chance has occurred in the sampling.

Practical applications of this theory are by people doing experiments and other samplings without knowledge of the population parameters. When they make confidence statements, they do not know whether they are right or wrong—they know only the probability selected.

EXAMPLE 3.9.1—Using the sample frequencies of table 3.8.1, test the hypothesis (known to be true) that the *t*-distribution is symmetrical in the sense that half of the population frequency is greater than zero. Ans. $\chi^2 = 1.22$.

EXAMPLE 3.9.2—From table 3.8.1 it is seen that $3 + 4 + 5 + 8 + 2 + 3 = 25$ samples have $|t| > 2.262$. Test the hypothesis that 5% of the population values are greater than $|2.262|$. Ans. $\chi^2 = 0.0124$

EXAMPLE 3.9.3—In table 3.8.1, accumulate the sample frequencies in both tails and compare their percentage values with those in the last column of the table.

EXAMPLE 3.9.4—During the fall of 1943, approximately one in each 1,000 city families of Iowa (cities are defined as having 2,500 inhabitants or more) was visited to learn the number of quarts of food canned. The average for 300 families was 165 quarts with standard deviation, 153 quarts. Calculate the 95% confidence limits. Ans. 165 ± 17 quarts.

EXAMPLE 3.9.5—The 1940 census reported 312,000 dwelling units (roughly the same as families) in Iowa cities. From the statistics of the foregoing example, estimate the number of quarts of food canned in Iowa cities in 1943. Ans. 51,500,000 quarts with 95% confidence limits, 46,200,000 and 56,800,000 quarts.

3.10—Use of frequency distributions for computing \bar{X} and s . In this chapter we have used frequency distributions formed by grouping the sample data into classes to give a picture of the way in which a variable is distributed in a population. A frequency distribution also provides a shortcut method of computing \bar{X} and s from a large sample. For this calculation, at least 12 classes are advisable, and for highly accurate work, at least 20 classes. The reason will be indicated presently.

After forming the classes and counting the frequency in each class, write down the class mark (center of the class) for each class. Normally, the class mark is found by noting the lower and the upper limits of the class, and taking the average of these two values. For instance, with data that are originally recorded to whole numbers, the class limits might be 0–9, 10–19, and so on. The class marks are 4.5, 14.5, and so on. Note that the marks are *not* 5, 15, etc., as we might hastily conclude.

The assumptions made in the shortcut computation are that the class mark is very close to the actual mean of the items in the class, and that these items are approximately evenly distributed throughout the class. These assumptions are likely to hold well in the high-frequency classes near the middle of the distribution. Caution is necessary if there are natural groupings in the scale of measurement. An instance was observed where the number of seed compartments in tomatoes was the variable, its values being confined to whole numbers and halves. However, halves occurred very infrequently. At first, the class intervals were chosen to extend from 2 up to but not including 3, etc., the class marks being written down as 2 1/2, 3 1/2, etc. Actually, the class means were almost at the lower boundaries, 2, 3, etc. This systematic error led to an overestimate of almost half a seed compartment in the mean. In this situation the actual class means should be computed and used as the class marks (see exercise 3.11.3).

The same problem can arise in the extreme classes in a frequency distribution. To revert to the example with intervals 0–9, 10–19, etc. and class marks taken as 4.5, 14.5, etc., we might notice that the lowest class contained six 0's, one 2, and one 6, so that the class mean is actually 1.0, whereas the class mark is 4.5. For accurate work the class mark for this class is taken as 1.0.

In the shortcut computation of \bar{X} and s , each item in the sample is replaced by the class mark for the class in which it lies. All values between 10 and 19 in the previous example are replaced by 14.5. The process is exactly the same as that of rounding to the nearest whole number, or the nearest 100. This rounding introduces an additional error into the data. The argument for having a relatively large number of classes is to keep this error small.

The remainder of this section discusses how much accuracy is lost owing to this rounding error. Let X represent any item in the sample and let X' be the corresponding class mark or rounded value. Then we may write

$$X' = X + e$$

where e is the rounding error. If I is the width of the class interval, the values e are assumed to be roughly evenly distributed over the range from $-I/2$ to $+I/2$. An important result from theory is that the variance of the sum of two independent variables is the sum of their variances. This gives

$$\sigma_{X'}^2 = \sigma_X^2 + \sigma_e^2$$

If e is uniformly distributed between $-I/2$ and $+I/2$, it is known from theory that its variance is $I^2/12$. Hence,

$$\sigma_{X'}^2 = \sigma_X^2 + I^2/12 = \sigma^2 + I^2/12,$$

since σ_X^2 is the original population variance σ^2 .

Consequently, when a value X is replaced by the corresponding class mark X' , the variance is increased by $I^2/12$ due to the rounding. The relative increase in variance is $I^2/12\sigma^2$. We would like this increase to be small.

Suppose that there are 12 classes in the frequency distribution. If the distribution is not far from normal, nearly all the frequency lies within a distance $\pm 3\sigma$ from μ . Since these classes cover a range of 6σ , I will be roughly $6\sigma/12 = \sigma/2$. Thus the relative increase in the variance of \bar{X} due to grouping is about $1/48$, or 2%. A further analysis, not presented here, shows that the computed s'^2 has a variance about 4% larger than that of the original s^2 (5). For ordinary work these small losses in accuracy to save time in computation are tolerable. For accurate work, the advice commonly given is that I should not exceed $\sigma/4$. This requires about 24 classes to cover the frequency distribution when the sample is large.

With a discrete variable, there is often no rounding and no loss of accuracy in using a frequency distribution to compute the sample mean and variance. For instance, in a study of accidents per week, the number of accidents might range only from 0 to 5. The six classes 0, 1, 2, 3, 4, 5, give a complete representation of the sample data without any rounding.

3.11—Computation of \bar{X} and s in large samples: example. The data in table 3.11.1 come from a sample of 533 weights of swine, arranged in 22 classes. The steps in the calculation of \bar{X} and s are given under the table.

A further simplification comes from *coding* the class marks, as shown in the third column. Place the 0 on the coded scale at or near the class mark that has the highest frequency. We chose this origin at $G = 170$ pounds. The classes above this class are coded, as 1, 2, 3, etc.; those

TABLE 3.11.1
 FREQUENCY DISTRIBUTION OF LIVE WEIGHTS OF 533 SWINE. COMPUTATION OF MEAN
 AND STANDARD DEVIATION. $I = 10$ POUNDS, $G = 170$ POUNDS

Class Mark, Pounds	Frequency f	Code Numbers U	Sum of Code Numbers fU	Squares fU^2
80	1	- 9	- 9	81
90	0	- 8	0	0
100	0	- 7	0	0
110	7	- 6	-42	252
120	18	- 5	-90	450
130	21	- 4	-84	336
140	22	- 3	-66	198
150	44	- 2	-88	176
160	67	- 1	-67	67
170	76	0	0	0
180	55	1	55	55
190	57	2	114	228
200	47	3	141	423
210	33	4	132	528
220	30	5	150	750
230	23	6	138	828
240	11	7	77	539
250	5	8	40	320
260	5	9	45	405
270	4	10	40	400
280	5	11	55	605
290	2	12	24	288
$n = 533$			$\Sigma fU = 565$	$\Sigma fU^2 = 6,929$
$\Sigma fU = 565$			$\Sigma fU^2 = 6,929$	
$I\bar{U} = 10(565/533)$			$(\Sigma fU)^2/n = (565)^2/533 = 598.92$	
$= 10.6$ pounds			$\Sigma u^2 = 6,330.08$	
$\bar{X} = G + I\bar{U}$			$s_U^2 = \Sigma u^2/(n-1) = 11.8986$	
$= 170 + 10.6$			$s_U = 3.45$	
$= 180.6$ pounds			$s = Is_U = (10)s_U = 34.5$ pounds	

below as -1 , -2 , -3 , etc. It is important to know the relation between your original and your coded class marks. If X (dropping the prime) is an original class mark and U is its coded value, this relation is

$$X = G + IU$$

where I is the width of the class interval (10 pounds in this example). To verify the rule, when U is -5 , what is X ? We have, $X = 170 + (10)(-5) = 120$, as appears in column 1.

In the computations we first find the sample mean and variance of U , namely \bar{U} and s_U . From the above relation we get

$$\bar{X} = G + I\bar{U}$$

and

$$s = s_X = Is_U$$

With these relations the steps given under table 3.11.1 are easily followed. With a computing machine the individual values fU^2 need not be written down. Their sum can be found by taking the sum of products of the column U with the column fU . The individual values fU are required: pay attention to their signs when adding them.

Note that s is 3.45 times the class interval I , so that the loss of accuracy due to the use of class marks is trivial.

Sheppard's Correction. From the theory presented in the previous section, a consequence is that s^2 , as computed in table 3.11.1, is an estimate of $\sigma^2 + I^2/12$, rather than of σ^2 itself. A correction introduced by W. F. Sheppard (6) is to subtract $I^2/12$ from the value of s^2 , in order to obtain a more nearly unbiased estimate of σ^2 . In this example, with $s^2 = 1,189.86$, the correction amounts to only $100/12$, or 8.33. The corrected value of s is 3.44 as against our computed 3.45. The correction is seldom substantial. The corrected value should not be used in a test of significance (7).

EXAMPLE 3.11.1—The data show the frequency distribution of the heights of 8,585 men, arranged in ten 2-in. classes. The number of classes is too small for accurate work, but gives an easy exercise. Compute \bar{X} and s , using a convenient coding. Ans. $\bar{X} = 67.53$ in., $s = 2.62$ in.

Class Mark (in.)	Frequency	Class Mark (in.)	Frequency
58	6	68	2,559
60	55	70	1,709
62	252	72	594
64	1,063	74	111
66	2,213	76	23

EXAMPLE 3.11.2—Apply Sheppard's correction and report the corrected s . Ans. 2.56 ins.

EXAMPLE 3.11.3—This baby example illustrates how the accuracy of the shortcut method improves when the class marks are the means of the items in the classes. The original data consist of the fourteen values: 0, 0, 10, 12, 14, 16, 20, 22, 24, 25, 29, 32, 34, 49. (i) Compute \bar{X} and s directly from these data. (ii) Form a frequency distribution with classes 0–9, 10–19, 20–29, 30–39, and 40–49. Compute \bar{X} and s from the conventional class marks, 4.5, 14.5, 24.5, 34.5, and 44.5. (iii) In the same frequency distribution, find the actual means of the items in each class, and use these means as the class marks. (Coding doesn't help here.) Ans. (i) $\bar{X} = 20.5$, $s = 13.4$. (ii) $\bar{X} = 21.6$, $s = 11.4$, both quite inaccurate. (iii) $\bar{X} = 20.5$, $s = 13.2$. Despite the rounding errors that contribute to this s , it is smaller than the original s in (i). This is an effect of sampling error in this small sample.

EXAMPLE 3.11 4—The yields in grams of 1,499 rows of wheat are recorded by Wiebe (9). They have been tabulated as follows:

Class Mark	Frequency	Class Mark	Frequency	Class Mark	Frequency
375	3	600	127	825	10
400	13	625	140	850	10
425	41	650	122	875	4
450	99	675	94	900	4
475	97	700	64	925	2
500	118	725	49	950	3
525	138	750	31	975	1
550	146	775	26	1,000	1
575	136	800	20		
				Total	1,499

Compute $\bar{X} = 587.74$ grams, and $s = 100.55$ grams. Are there enough classes in this distribution?

3.12—Tests of normality. Since many of the standard statistical techniques are based on the assumption of normality, methods for judging the normality of a set of data are of interest. In this and in the following sections, three tests will be illustrated from the frequency distribution of means of samples of 100 drawn from the population of city sizes in section 2.12 (p. 51). The histogram of this frequency distribution, shown in the bottom part of figure 2.12.2, p. 55, gave the impression that a normal distribution would not be a good fit. We can now verify this impression in a quantitative manner.

In the first test, often called the χ^2 *goodness of fit test*, the data are grouped into classes to form a frequency distribution and the sample mean \bar{X} and standard deviation s are calculated. From these quantities, a normal distribution is fitted and the expected frequencies in each class are obtained as described in section 3.4 (p. 70). Table 3.12.1 presents the observed frequencies f_i and the expected frequencies F_i .

For each class, compute and record the quantity

$$(f_i - F_i)^2 / F_i = (\text{Obs.} - \text{Exp.})^2 / \text{Exp.}$$

The test criterion is

$$\chi^2 = \sum (f_i - F_i)^2 / F_i$$

summed over the classes. If the data actually come from a normal distribution, this quantity follows approximately the theoretical χ^2 distribution with $(k - 3) d.f.$, where k is the number of classes used in computing χ^2 . If the data come from some other distribution, the observed f_i will tend to agree poorly with the values of F_i that are expected on the assumption of normality, and the computed χ^2 becomes large. Consequently, large values of χ^2 cause rejection of the hypothesis of normality.

TABLE 3.12.1
CALCULATION OF THE GOODNESS OF FIT χ^2 FOR THE DISTRIBUTION OF MEANS OF
SAMPLES OF 100 CITY SIZES

Class Limits (1,000's)	Frequencies		
	Obs f_i	Exp F_i	$(f_i - F_i)^2/F_i$
Under 129	9	20.30	6.29
130-139	35	30.80	0.57
140-149	68	55.70	2.72
150-159	94	80.65	2.21
160-169	90	93.55	0.13
170-179	76	87.00	1.39
180-189	62	64.80	0.12
190-199	28	38.70	2.96
200-209	27	18.45	3.85
210-219	4	7.10	1.35
220-229	5	2.20	6.04
230-239	1	0.50	
240-	1	0.15	
Total	500	500.00	27.63

$$\chi^2 = 27.63, df = 11 - 3 = 8 \quad P < 0.005$$

The theorem that this quantity follows the theoretical distribution of χ^2 when the null hypothesis holds and that the degrees of freedom are $(k - 3)$ requires advanced methods of proof. The subtracted number 3 in the $d.f.$ may be thought of as the number of ways in which the observed and expected frequencies have been forced to agree in the process of fitting the normal distribution. The numbers f_i and F_i both add to 500 and the sets agree in the values of \bar{X} and s that they give.

The theorem also requires that the expected numbers not be too small. Small expectations are likely to occur only in the extreme classes. A working rule (10) is that the two extreme expectations may each be as low as 1, provided that most of the other expected values exceed 5. In table 3.12.1, small expectations occur in the three highest classes. In this event, classes are combined to give an expectation of at least one. The three highest classes give a combined f_i of 7 and F_i of 2.85. The contribution to χ^2 is $(4.15)^2/2.85 = 6.04$.

For these data, $k = 11$ after combination, so that $\chi^2 = 27.63$ has 8 $d.f.$ Reference to table A 5 shows that the hypothesis of normality is rejected at the 0.5% level, the most extreme level given in this table.

The χ^2 test may be described as a non-specific test, in that the test criterion is directed against no particular type of departure from normality. Examples occur in which the data are noticeably skew, although the χ^2 test does not reject the null hypothesis. An alternative test that is designed to detect skewness is often used as a supplement to the χ^2 test.

3.13—A test of skewness. A measure of the amount of skewness in a population is given by the average value of $(X - \mu)^3$, taken over the population. This quantity is called the *third moment about the mean*. If low values of X are bunched close to the mean μ but high values extend far above the mean, this measure will be positive, since the large positive contributions $(X - \mu)^3$ when X exceeds μ will predominate over the smaller negative contributions $(X - \mu)^3$ obtained when X is less than μ . Populations with negative skewness, in which the lower tail is the extended one, are also encountered. To render this measure independent of the scale on which the data are recorded, it is divided by σ^3 . The resulting *coefficient of skewness* is denoted sometimes by $\sqrt{\beta_1}$ and sometimes by γ_1 .

The sample estimate of this coefficient is denoted by $\sqrt{b_1}$ or g_1 . We compute

$$\begin{aligned} m_3 &= \Sigma(X - \bar{X})^3/n \\ m_2 &= \Sigma(X - \bar{X})^2/n \end{aligned}$$

and take

$$\sqrt{b_1} = g_1 = m_3/(m_2\sqrt{m_2})$$

Note that in computing m_2 , the sample variance, we have divided by n instead of our customary $(n - 1)$. This makes subsequent calculations slightly easier.

The calculations are illustrated for the means of city sizes in table 3.13.1. Coding is worthwhile. Since $\sqrt{b_1}$ is dimensionless, the whole calculation can be done in the coded scale, with no need to decode. Having chosen coded values U , write down their squares and cubes (paying attention to signs). The U^4 values are not needed in this section. Form the sums of products with the f 's as indicated, and divide each sum by n to give the quantities h_1 , h_2 , h_3 . Carry two extra decimal places in the h 's. The moments m_2 and m_3 are then obtained from the algebraic identities given under the table. Finally, we obtain $\sqrt{b_1} = 0.4707$.

If the sample comes from a normal population, $\sqrt{b_1}$ is approximately normally distributed with mean zero and *S.D.* $\sqrt{(6/n)}$, or in this case $\sqrt{(6/500)} = 0.110$. Since $\sqrt{b_1}$ is over 4 times its *S.D.*, the positive skewness is confirmed. The assumption that $\sqrt{b_1}$ is normally distributed is accurate enough for this test if n exceeds 150. For sample sizes between 25 and 200, the *one-tailed* 5% and 1% significance levels of $\sqrt{b_1}$, computed from a more accurate approximation, are given in table A 6.

3.14—Tests for kurtosis. A further type of departure from normality is called *kurtosis*. In a population, a measure of kurtosis is the average value of $(X - \mu)^4$, divided by σ^4 . For the normal distribution, this ratio has the value 3. If the ratio exceeds 3, there is usually an excess of values near the mean and far from it, with a corresponding depletion of the flanks of the distribution curve. This is the manner in which the *t*-distribution

TABLE 3.13.1
COMPUTATIONS FOR TESTS OF SKEWNESS AND KURTOSIS

Lower Class Limit	f	U	U^2	U^3	U^4
120-	9	-4	16	-64	256
130-	35	-3	9	-27	81
140-	68	-2	4	-8	16
150-	94	-1	1	-1	1
160-	90	0	0	0	0
170-	76	1	1	1	1
180-	62	2	4	8	16
190-	28	3	9	27	81
200-	27	4	16	64	256
210-	4	5	25	125	625
220-	5	6	36	216	1,296
230-	1	7	49	343	2,401
240-	1	8	64	512	4,096

$n = 500$

Test of skewness

$$\Sigma fU = +86 \quad h_1 = \Sigma fU/n = +0.172$$

$$\Sigma fU^2 = 2,226 \quad h_2 = \Sigma fU^2/n = 4.452$$

$$\Sigma fU^3 = +3,332 \quad h_3 = \Sigma fU^3/n = +6.664$$

$$m_2 = h_2 - h_1^2 = 4.4224$$

$$m_3 = h_3 - 3h_1h_2 + 2h_1^3 = 4.3770$$

$$\sqrt{b_1} = m_3/m_2\sqrt{m_2} = 4.3770/(4.4224)\sqrt{4.4224} = 0.4707$$

Test of kurtosis

$$\Sigma fU^4 = 32,046 \quad h_4 = \Sigma fU^4/n = 64.092$$

$$m_4 = h_4 - 4h_1h_3 + 6h_1^2h_2 - 3h_1^4 = 60.2948$$

$$b_2 = m_4/m_2^2 = 60.2948/(4.4224)^2 = 3.083$$

departs from the normal. Ratios less than 3 result from curves that have a flatter top than the normal.

A sample estimate of the amount of kurtosis is given by

$$g_2 = b_2 - 3 = (m_4/m_2^2) - 3,$$

where

$$m_4 = \Sigma (X - \bar{X})^4/n$$

is the fourth moment of the sample about its mean. Notice that the normal distribution value 3 has been subtracted, with the result that peaked distributions show positive kurtosis and flat-topped distributions show negative kurtosis.

The shortcut computation of m_4 and b_2 from the coded values U is shown under table 3.13.1. For this sample, $g_2 = b_2 - 3$ has the value +0.083. In very large samples from the normal distribution, g_2 is normally distributed with mean 0 and $S.D. \sqrt{(24/n)} = 0.219$, since n is 500. The sample value of g_2 is much smaller than its standard error, so that the amount of kurtosis in the population appears to be trivial.

Unfortunately, the distribution of g_2 does not approach the normal closely until the sample size is over 1,000. For sample sizes between 200 and 1,000, table A 6 contains better approximations to the 5% and 1% significance levels. Since the distribution of g_2 is skew, the two tails are shown separately. For $n = 500$, the upper 5% value of g_2 is $+0.37$, much greater than the value 0.083 found in this sample.

For sample sizes less than 200, no tables of the significance levels of g_2 are at present available. R. C. Geary (11) developed an alternative test criterion for kurtosis,

$$a = (\text{mean deviation})/(\text{standard deviation}) \\ = \Sigma |X - \bar{X}| / n\sqrt{m_2},$$

and tabulated its significance levels for sample sizes down to $n = 11$. If X is a normal deviate, the value of a when computed for the whole population is 0.7979. Positive kurtosis produces higher values, and negative kurtosis lower values of a . When applied to the same data, a and g_2 usually agree well in their verdicts. The advantages of a are that tables are available for smaller sample sizes and that a is easier to compute.

An identity simplifies the calculation of the numerator of a . This will be illustrated for the coded scale in table 3.13.1. Let

$$\begin{aligned} \Sigma' &= \text{sum of all observations that exceed } \bar{U} \\ n' &= \text{number of observations that exceed } \bar{U} \\ \Sigma |U - \bar{U}| &= 2(\Sigma' - n'\bar{U}) \end{aligned}$$

Since $\bar{U} = 0.172$, all observations in the classes with $U = 1$ or more exceed \bar{U} . This gives $\Sigma' = 457$, $n' = 204$. Hence,

$$\Sigma |U - \bar{U}| = 2\{457 - (204)(0.172)\} = 843.82$$

Since $m_2 = 4.4224$, we have

$$a = (843.82)/(500)\sqrt{4.4224} = 0.802$$

This is little greater than the value 0.7979 for the normal distribution, in agreement with the result given by g_2 . For $n = 500$ the upper 5% level of a is about 0.814.

3.15—Effects of skewness and kurtosis. In samples from non-normal populations, the quantities g_1 and g_2 are useful as estimates of the corresponding population values γ_1 and γ_2 , which characterize the common types of non-normality. K. Pearson produced a family of theoretical non-normal curves intended to simulate the shapes of frequency distributions having any specified values of γ_1 and γ_2 , provided that the non-normality was not too extreme.

The quantities γ_1 and γ_2 have also been useful in studying the distributions of \bar{X} and s^2 when the original population is non-normal. Two results will be quoted. For the distribution of \bar{X} in random samples of size n ,

$$\gamma_1(\bar{X}) = \gamma_1/\sqrt{n} : \gamma_2(\bar{X}) = \gamma_2/n$$

Thus, in the distribution of \bar{X} , the measures of skewness and kurtosis both go to zero when the sample size increases, as would be expected from the Central Limit Theorem. Since the kurtosis is damped much faster than the skewness, it is not surprising that in our sample means g_1 was substantial but g_2 small.

Secondly, the exact variance of s^2 with f degrees of freedom is known to be

$$V(s^2) = \frac{2\sigma^4}{f} \left\{ 1 + \frac{f}{f+1} \cdot \frac{\gamma_2}{2} \right\}$$

The factor outside the brackets is the variance of s^2 in samples from a normal population. The term inside the brackets is the factor by which the normal variance is multiplied when the population is non-normal. For example, if the measure of kurtosis, γ_2 , is 1, the variance of s^2 is about 1.5 times as large as it is in a normal population. With $\gamma_2 = 2$, the variance of s^2 is about twice as large as in a normal population. These results show that the distribution of s^2 is sensitive to amounts of kurtosis that may pass unnoticed in handling the data.

EXAMPLE 3.15.1—In table 3.2.2, compute $g_1 = -0.0139$ and $g_2 = 0.0460$, showing that the distribution is practically normal in these respects.

EXAMPLE 3.15.2—In table 3.5.2 is the sampling distribution of 511 standard deviations. Calculate $g_1 = 0.3074$ with standard error 0.108. As expected, this indicates that the distribution is positively skew.

EXAMPLE 3.15.3—The 511 values of t discussed in section 3.8 were distributed as follows:

Class Mark	f	Class Mark	f	Class Mark	f	Class Mark	f
-3.13	3	-1.13	29	0.87	31	2.87	1
-2.88	5	-0.88	35	1.12	23	3.12	1
-2.63	1	-0.63	38	1.37	17	3.37	2
-2.38	3	-0.38	40	1.62	11	3.62	0
-2.13	6	-0.13	52	1.87	8	3.87	0
-1.88	12	0.12	57	2.12	10	4.12	0
-1.63	21	0.37	43	2.37	6	4.37	1
-1.38	16	0.62	37	2.62	2		
						Total	511

The highly significant value of $g_2 = 0.5340$ shows that the frequencies near the mode and in the tails are greater than in the normal distribution, those in the flanks being less. This was expected. But $g_1 = 0.1356$ is non-significant, which is also expected because the theoretical distribution of t is symmetrical.

REFERENCES

1. RAND CORPORATION. *A Million Random Digits With 100,000 Normal Deviates*. Free Press, Glencoe, Ill. (1955).

90 Chapter 3: Experimental Sampling From a Normal Population

2. P. C. MAHALANOBIS, *et al.* *Sankhyā*, 1:1 (1934).
3. E. S. PEARSON and H. O. HARTLEY. *Biometrika Tables for Statisticians*, Vol I. Cambridge University Press (1954).
4. N. V. SMIRNOY. *Tables for the Distribution and Density Functions of t-distribution*. Pergamon Press, New York (1961).
5. R. A. FISHER. *Phil. Trans.*, A, 222:309 (1921).
6. W. F. SHEPPARD. *Proc. Lond. Math. Soc.*, 29:353 (1898).
7. R. A. FISHER. *Statistical Methods for Research Workers*, 13th ed. Oliver and Boyd, Edinburgh (1958).
8. E. W. LINDSTROM. *Amer. Nat.*, 49:311 (1935).
9. G. A. WIEBE. *J. Agric. Res*, 50:331 (1935).
10. W. G. COCHRAN. *Biometrics*, 10:420 (1954).
11. R. C. GEARY. *Biometrika*, 28:295 (1936).

The comparison of two samples

4.1—Estimates and tests of differences. Investigations are often designed to discover and evaluate *differences* between effects rather than the effects themselves. It is the difference between the amounts learned under two methods of teaching, the difference between the lengths of life of two types of glassware or the difference between the degrees of relief reported from two pain-relieving drugs that is wanted. In this chapter we consider the simplest investigation of this type, in which two groups or two procedures are compared. In experimentation, these procedures are often called the *treatments*. Such a study may be conducted in two ways.

Paired samples. Pairs of *similar* individuals or things are selected. One treatment is applied to one member of each pair, the other treatment to the second member. The members of a pair may be two students of similar ability; two patients of the same age and sex who have just undergone the same type of operation; or two male mice from the same litter. A common application occurs in *self-pairing* in which a single individual is measured on two occasions. For example, the blood pressure of a subject might be measured before and after heavy exercise. For any pair, the difference between the measurements given by the two members is an estimate of the difference in the effects of the two treatments or procedures.

With only a single pair it is impossible to say whether the difference in behavior is to be attributed to the difference in treatment, to the natural variability of the individuals, or partly to both. There must be a number of pairs. The data to be analyzed consist of a sample of n differences in measurement.

Independent samples. This case, which is commoner, arises whenever we wish to compare the means of two populations and have drawn a sample from each quite independently. We might have a sample of men aged 50–55 and one of men aged 30–35, in order to compare the amounts spent on life insurance. Or we might have a sample of high school seniors from rural schools and one from urban schools, in order to compare their knowledge of current affairs as judged by a special examination on

this subject. Independent samples are widely used in experimentation when no suitable basis for pairing exists, as, for example, in comparing the lengths of life of two types of drinking glass under the ordinary conditions of restaurant use.

4.2—A simulated paired experiment. Eight pairs of random normal deviates were drawn from a table of random normal deviates. The first member of each pair represents the result produced by a Standard procedure, while the second member is the result produced by a New procedure that is being compared with the Standard. The eight differences, New-St., are shown in the Column headed Case I in table 4.2.1.

TABLE 4.2.1
A SIMULATED PAIRED EXPERIMENT

Pairs	CASE I New-St. (D_i)	CASE II New-St. (D_i)	CASE III New-St. (D_i)
1	+3.2	+13.2	+4.2
2	-1.7	+8.3	-0.7
3	+0.8	+10.8	+1.8
4	-0.3	+9.7	+0.7
5	+0.5	+10.1	+1.5
6	+1.2	+11.2	+2.2
7	-1.1	+8.9	-0.1
8	-0.4	+9.6	+0.6
Mean (\bar{D})	+0.28	+10.28	+1.28
s_D	1.527	1.527	1.527
s_D	0.540	0.540	0.540

Since the results for the New and Standard procedures were drawn from the same normal population, Case I simulates a situation in which there is no difference in effect between the two procedures. The observed differences represent the natural variability that is always present in experiments. It is obvious on inspection that the eight differences do not indicate any superiority of the New procedure. Four of the differences are + and 4 are - and the mean difference is small.

The results in Case II were obtained from those in Case I by adding +10 to every figure, to represent a situation in which the New procedure is actually 10 units better than the Standard. On looking at the data, most investigators would reach the judgment that the superiority of the New procedure is definitely established, and would probably conclude that the average advantage in favor of it is not far from 10 units.

Case III is more puzzling. We added +1 to every figure in Case I, so that the New procedure gives a small gain over the Standard. The New procedure wins 6 times out of the 8 trials, and some workers might conclude that the results confirm the superiority of the New procedure.

Others might disagree. They might point out that it is not too unusual for a fair coin to show heads in 6 tosses out of 8, and that the individual results range from an advantage of 0.7 units for the Standard to an advantage of 4.2 units for the New procedure. They would argue that the results are inconclusive. We shall see what verdicts are suggested by the statistical analyses in these three cases.

The data also illustrate the assumptions made in the analysis of a paired trial. The differences D_i in the individual pairs are assumed to be distributed about a mean μ_D , which represents the average difference in the effects of the two treatments over the population of which these pairs are a random sample. The deviations $D_i - \mu_D$ may be due to various causes, in particular to inherent differences between the members of the pair and to any errors of measurement to which the measuring instruments are subject. Another source of this variation is that a treatment may actually have different effects on different members of the population. A lotion for the relief of muscular pains may be more successful with some types of pain than with others. The adage: "One man's meat is another man's poison" expresses this variability in extreme form. For many applications it is important to study the extent to which the effect of a treatment varies from one member of the population to another. This requires a more elaborate analysis, and usually a more complex experiment, than we are discussing at present. In the simple paired trial we compare only the *average* effects of the two treatments or procedures over the population.

In the analysis, the deviations $D_i - \mu_D$ are assumed to be normally and independently distributed with population mean zero. The consequences of failures in these assumptions are discussed in chapter 11.

When these assumptions hold, the sample mean difference \bar{D} is normally distributed about μ_D with standard deviation or standard error σ_D/\sqrt{n} , where σ_D is the *S.D.* of the population of differences. The value of σ_D is seldom known, but the sample furnishes an estimate

$$s_D = \sqrt{\frac{\sum (D_i - \bar{D})^2}{n-1}} = \sqrt{\frac{\sum D_i^2 - (\sum D_i)^2/n}{n-1}}$$

Hence, $s_D = \sigma_D/\sqrt{n}$ is an estimate of σ_D , based on $(n-1)$ *d.f.*

The important consequence of these results is that the quantity

$$t = (\bar{D} - \mu_D)/s_D$$

follows Student's *t*-distribution with $(n-1)$ *d.f.*, where n is the number of pairs. The *t*-distribution may be used to test the null hypothesis that $\mu_D = 0$, or to compute a confidence interval for μ_D .

Test of significance The test will be applied first to the doubtful Case III. The values of s_D and \bar{s}_D are shown at the foot of table 4.2.1. Note that these are exactly the same in all three cases, since the addition of a constant μ_D to all the D_i does not affect the deviations $(D_i - \bar{D})$. For Case III we have

$$t = \bar{D}/s_{\bar{D}} = 1.28/0.540 = 2.370$$

With 7 *d.f.*, table A 4 shows that the 5% level of *t* in a two-tailed test is 2.365. The observed mean difference just reaches the 5% level, so that the data point to a superiority of the new treatment.

In Case II, $t = 10.28/0.540 = 19.04$. This value lies far beyond even the 0.1% level (5.405) in table A 4. We might report: " $P < 0.001$."

In Case I, $t = 0.28/0.540 = 0.519$. From table A 4, an absolute value of $t = 0.711$ is exceeded 50% of the time in sampling from a population with $\mu_D = 0$. The test provides no evidence on which to reject the null hypothesis in Case I. To sum up, the tests confirm the judgment of the preliminary inspection in all three cases.

Confidence interval. From the formula given in section 2.16, the 95% confidence interval for μ_D is

$$\bar{D} \pm t_{0.05} s_{\bar{D}} = \bar{D} \pm (2.365)(0.540) = \bar{D} \pm 1.28$$

In the simulated example the limits are as follows.

Case I :	- 1.00 to 1.56
Case II :	9.00 to 11.56
Case III:	0.00 to 2.56

As always happens, the 95% confidence limits agree with the verdict given by the 5% tests of significance. Either technique may be used.

4.3—Example of a paired experiment. The preceding examples illustrate the assumptions and formulas used in the analysis of a paired set of data, but do not bring out the purpose of the pairing. Youden and Beale (1) wished to find out if two preparations of a virus would produce different effects on tobacco plants. The method employed was to rub half a leaf of a tobacco plant with cheesecloth soaked in one preparation of the virus extract, then to rub the second half similarly with the second extract. The measurement of potency was the number of local lesions appearing on the half leaf: these lesions appear as small dark rings that are easily counted. The data in table 4.3.1 are taken from leaf number 2 on each of 8 plants. The steps in the analysis are exactly the same as in the preceding. We have, however, presented the deviations of the differences from their mean, $d_i = D_i - \bar{D}$, and obtained the sum of squares of deviations directly instead of by the shortcut formula.

For a test of the null hypothesis that the two preparations produce on the average the same number of lesions, we compute

$$t = \frac{\bar{D}}{s_{\bar{D}}} = \frac{4}{1.52} = 2.63, \quad d.f. = n - 1 = 7$$

From table A 4, the significance probability is about 0.04, and the null hypothesis is rejected. We conclude that in the population the second preparation produces fewer lesions than the first. From this result we

TABLE 4.3.1
NUMBER OF LESIONS ON HALVES OF EIGHT TOBACCO LEAVES*

Pair No.	Preparation 1 X_1	Preparation 2 X_2	Difference $D = X_1 - X_2$	Deviation $d = D - \bar{D}$	Squared Deviation d^2
1	31	18	13	9	81
2	20	17	3	-1	1
3	18	14	4	0	0
4	17	11	6	2	4
5	9	10	-1	-5	25
6	8	7	1	-3	9
7	10	5	5	1	1
8	7	6	1	-3	9
Total	120	88	32	0	130
Mean	15	11	$\bar{D} = 4$		$s_D^2 = 18.57$

$$s_D^2 = 18.57/8 = 2.32, s_D = 1.52 \text{ lesions}$$

* Slightly changed to make calculation easier.

would expect that both the 95% confidence limits for μ_D will be positive. Since $t_{0.05} s_D = (2.365)(1.52) = 3.69$, the 95% limits are +0.4 and +7.6 lesions per leaf.

In this experiment the leaf constitutes the pair. This choice was made as a result of earlier studies in which a single preparation was rubbed on a large number of leaves, the lesions found on each half-leaf being counted. In a new type of work, a preliminary study of this kind can be highly useful. Since every half-leaf was treated in the same way, the variations found in the numbers of lesions per half leaf represent the natural variability of the experimental material. From the data, the investigator can estimate the population standard deviation, from which he can in turn estimate the size of sample needed to ensure a specified degree of precision in the sample averages. He can also look for a good method of forming pairs. Such a study is sometimes called a *uniformity trial*, because the treatment is uniform, although a *variability trial* might be a better name.

Youden and Beale found that the two halves of the same leaf were good partners, since they tended to give similar numbers of lesions. An indication of this fact is evident in table 4.3.1, where the pairs are arranged in descending order of total numbers of lesions per leaf. Notice that with two minor exceptions, this descending order shows up in each preparation. If one member of a pair is high, so is the other; if one is low, so is the other. The numbers on the two halves of a leaf are said to be *positively correlated*. Because of this correlation, the differences between the two halves tend to be mostly small, and therefore less likely to mask or conceal an imposed difference due to a difference in treatments.

96 Chapter 4: The Comparison of Two Samples

EXAMPLE 4.3.1—L. C. Grove (2) determined the sample mean numbers of florets produced by seven pairs of plots of Excellence gladiolus, one plot of each pair planted with high (first year) corms, the other with low (second-year or older) corms (A corm is an underground propagating stem). The plot means were as follows

Corm		Florets					
High	11.2	13.3	12.8	13.7	12.2	11.9	12.1
Low	14.6	12.6	15.0	15.6	12.7	12.0	13.1

Calculate the sample mean difference. **Ans.** 1.2 florets. In the population of such differences, test the null hypothesis $\mu_D = 0$. **Ans.** $P = 0.06$, approximately.

EXAMPLE 4.3.2—Samples of blood were taken from each of 8 patients. In each sample, the serum albumen content of the blood was determined by each of two laboratory methods A and B. The objective was to discover whether there was a consistent difference in the amount of serum albumen found by the two methods. The 8 differences (A-B) were as follows: 0.6, 0.7, 0.8, 0.9, 0.3, 0.5, -0.5, 1.3, the units being gm per 100 ml. Compute t to test the null hypothesis (H_0) that the population mean of these differences is zero, and report the approximate value of your significance probability. What is the conclusion? **Ans.** $t = 2.511$ with 7 d.f. P between 0.05 and 0.025. Method A has a systematic tendency to give higher values.

EXAMPLE 4.3.3—Mitchell, Burroughs, and Beadles (3) computed the biological values of proteins from raw peanuts (P) and roasted peanuts (R) as determined in an experiment with 10 pairs of rats. The pairs of data P, R are as follows: 61, 55, 60, 54, 56, 47, 63, 59, 56, 51, 63, 61, 59, 57, 56, 54, 44, 63, 61, 58. Compute the sample mean difference, t , and the sample standard deviation of the differences, 7.72 units. Since $t = 0.82$, over 40% of similar samples from a population with $\mu_D = 0$ would be expected to have larger t -values.

Note: 9 of the 10 differences, $P - R$, are positive. One would like some information about the next-to-the-last pair 44, 63. The first member seems abnormal. While unusual individuals like this do occur in the most carefully conducted trials, their appearance demands immediate investigation. Doubtless an error in recording or computation was searched for but not found. What should be done about such aberrant observations is a moot question; their occurrence detracts from one's confidence in the experiment.

EXAMPLE 4.3.4—A man starting work in a new town has two routes A and B by which he may drive home. He conducts an experiment to find out which route is quicker. Since traffic is unusually heavy on Mondays and Fridays but does not seem to vary much from week to week, he selects the day of the week as the basis for pairing. The test lasts four weeks. On the first Monday, he tosses a coin to decide whether to drive by route A or B. On the second Monday, he drives by the other route. On the third Monday, he again tosses a coin, using the other route on the fourth Monday, and similarly for the other days of the week. The times taken, in minutes, were as follows:

	M1	M2	Tu1	Tu2	W1	W2	Th1	Th2	F1	F2
A	28.7	26.2	24.8	25.3	25.1	23.9	26.1	25.8	30.3	31.4
B	25.4	25.8	24.9	25.0	23.9	23.3	26.6	24.8	28.8	30.3

(i) Treating the data as consisting of 10 pairs, test whether there seems to be any real difference in average driving times between A and B. (ii) Compute 95% confidence limits for the population mean difference. What would you regard as the population in this trial? (iii) By eye inspection of the results, does the pairing look effective? (iv) Suppose that on the last Friday (F2) there had been a fire on route B, so that the time taken to get home was 48

minutes Would you recommend rejecting this pair from the analysis? Give your reason
 Ans (i) $t = 2.651$, with 9 *df* P about 0.03 Method B seems definitely quicker, (ii) 0.12 to 1.63 mins There really isn't much difference (iii) Highly effective

4.4—Conditions for pairing. The objective of pairing is to increase the precision of the comparison of the two procedures. Identical twins are natural pairs. Litter mates of the same sex are often paired successfully, because they usually behave more nearly alike than do animals less closely related. If the measurement at the end of the experiment is the subject's ability to perform some task (e.g., to do well in an exam), subjects similar in natural ability and previous training for this task should be paired. Often the subjects are tested at the beginning of the trial to provide information for forming pairs. Similarly, in experiments that compare two methods of treating sick persons, patients whose prognosis appears about the same at the beginning of the trial should be paired if feasible.

The variable on which we pair should predict accurately the performance of the subjects *on the measurement by which the effects of the treatments are to be judged*. Little will be gained by pairing students on their I.Q.'s if I.Q. is not closely related to ability to perform the particular task that is being measured in the experiment.

Self-pairing is highly effective when an individual's performance is consistent on different occasions, but yet exhibits wide variation when comparisons are made from one individual to another. If two methods of conducting a chemical extraction are being compared, the pair is likely to be a sample of the original raw material which is thoroughly mixed and divided into two parts.

Environmental variation often calls for pairing. Two treatments should be laid down side by side in the field or on the greenhouse bench in order to avoid the effects of unnecessary differences in soil, moisture, temperature, etc. Two plots or pots next to each other usually respond more nearly alike than do those at a distance. As a final illustration, sometimes the measuring process is lengthy and at least partly subjective, as in certain psychiatric studies. If several judges must be used to make the measurements for comparing two treatments A and B, each scoring a different group of patients, an obvious precaution is to ensure that each judge scores as many A patients as B patients. Even if the patients were not originally paired, they could be paired for assignment to judges.

Before an experiment has been conducted, it is of course not possible to foretell how effective a proposed pairing will be in increasing precision. However, from the results of a paired experiment, its precision may be compared with that of the corresponding unpaired experiment (section 4.11).

4.5—Tests of other null hypotheses about μ The null hypothesis $\mu_D = 0$ is not the only one that is useful, and the *alternative* may be $\mu_D > 0$ instead of $\mu_D \neq 0$. Illustrations are found in a Boone County survey of

corn borer effects. On 14 farms, the effect of spraying was evaluated by measuring the corn yield from both sprayed and unsprayed strips in each field. The data are recorded in table 4.5.1. The sample mean difference is 4.7 bu./acre with $s_D = 6.48$ bu./acre and $s_D/\sqrt{14} = 1.73$ bu./acre.

A one-tailed t -test. It had already been established that the spray, at the concentration used, could not decrease yield. If there is a decrease, as in the first field, it must be attributed to causes other than the spray, or to sampling variation. Consequently if μ_D is not zero then it must be greater

TABLE 4.5.1
YIELDS OF CORN (BUSHELS PER ACRE) IN SPRAYED AND UNSPRAYED STRIPS OF 14 FIELDS
Boone County, Iowa, 1950

Sprayed	64.3	78.1	93.0	80.7	89.0	79.9	90.6	102.4
Unsprayed	70.0	74.4	86.6	79.2	84.7	75.1	87.3	98.8
Difference	-5.7	3.7	6.4	1.5	4.3	4.8	3.3	3.6

70.7	106.1	107.4	74.0	72.6	69.5
70.2	101.1	83.4	65.2	68.1	68.4
0.5	5.0	24.0	8.8	4.5	1.1

than zero. The objective of this experiment was to test $H_0: \mu_D = 0$ with $H_A: \mu_D > 0$. As before,

$$t = \frac{4.7 - 0}{1.73} = 2.72, df = 13$$

To make a one-tailed test with table A 4, locate the sample value of t and use half of the probability indicated.

Applying this rule to the $t = 2.72$ above, P is slightly less than 0.02/2; the null hypothesis is rejected at $P < 0.01$. Evidently spraying did decrease corn borer damage, resulting in increased yields in Boone County in 1950.

Test of a non-zero μ . This same Boone County experiment may be cited to illustrate the use of a null hypothesis different from $\mu_D = 0$. This experiment might have had as its objective the test of the null hypothesis, "The cost of spraying is equal to the gain from increased yield." To evaluate costs, the fee of commercial sprayers was \$3 per acre and the 1950 crop was sold at about \$1.50 per bushel. So 2 bushels per acre would pay for the spraying. This test would be $H_0: \mu_D = 2$ bu./acre, $H_A: \mu_D \neq 2$ bu. acre, resulting in

$$t = \frac{4.7 - 2.0}{1.73} = 1.56, df = 13$$

The two-tailed probability is about $P = 0.15$, and the null hypothesis would presumably not be rejected. The verdict of the test is inconclusive: it provides no strong evidence that the farmers will either gain or lose by spraying.

One-tailed test of a non-zero μ . It is possible that $H_0: \mu_D = 2$ bu./acre might be tested with $H_A: \mu_D > 2$ bu./acre; that is, the alternative hypothesis might be put in the form of a slogan, "It pays to spray." If this were done, $t = 1.56$ would be associated with $P = 0.15/2 = 0.075$, not significant. But the implication of this one-sided test is that H_0 would be accepted no matter how far the sample mean might fall short of 2 bu./acre. It is the two tailed test which is appropriate here.

This point is stressed for the reason that some people use the one-sided test because, as a man said, "I am not interested in the other alternative." A one-tailed test of $H_0: \mu_D = \mu_0$ against $H_A: \mu_D > \mu_0$ should be used only if we know enough about the nature of the process being studied to be certain that μ_D could not be less than μ_0 .

In considering the profitability of spraying, it is more informative to treat the statistical problem as one of estimation than as one of testing hypotheses. Since the mean difference in yield between sprayed and unsprayed strips is 4.7 bu. per acre, the sample estimate of the profit per acre due to spraying is 2.7 bu. We can compute confidence limits for the average profit per acre over a population of fields of which this is a random sample. For 90% limits we add and subtract $t_{0.10} s_D = (1.771)(1.73) = 3.1$ bu. Thus if the farmers are willing to take a 1-in-10 chance that the sample estimate was not exceptionally poor, they learn that the average profit per acre lies somewhere between -0.4 bu. and +5.8 bu. These limits are unfortunately rather wide for a practical decision: a larger sample size would be necessary to narrow the limits. They do indicate, however, that although there is the possibility of a small loss, there is also the possibility of a substantial profit. The 95% limits, -1.0 bu. and +6.4 bu., tell much the same story.

EXAMPLE 4.5.1—In an investigation of the effect of feeding 10 mcg. of vitamin B_{12} per pound of ration to growing swine (4), 8 lots (each with 6 pigs) were fed in pairs. The pairs were distinguished by being fed different levels of aureomycin, an antibiotic which did not interact with the vitamin; that is, the differences were not affected by the aureomycin. The average daily gains (to about 200 lbs. live weight) are summarized as follows:

Ration	Pairs of Lots							
	1	2	3	4	5	6	7	8
With B_{12}	1.60	1.68	1.75	1.64	1.75	1.79	1.78	1.77
Without B_{12}	1.56	1.52	1.52	1.49	1.59	1.56	1.60	1.56
Difference, D	0.04	0.16	0.23	0.15	0.16	0.23	0.18	0.21

For the differences, calculate the statistics, $\bar{D} = 0.170$ lb./day and $s_D = 0.0217$ lb./day.

EXAMPLE 4.5.2—It is known that the addition of small amounts of the vitamin can not decrease the rate of growth. While it is fairly obvious that \bar{D} will be found significantly different from zero, the differences being all positive and, with one exception, fairly consistent, you may be interested in evaluating t . Ans. 7.83, far beyond the 0.01 level in the table. The appropriate alternative hypothesis is $\mu > 0$.

EXAMPLE 4.5.3—The effect of B_{12} seems to be a stimulation of the metabolic processes including appetite. The pigs eat more and grow faster. In the experiment above, the cost of the additional amount of feed eaten, including that of the vitamin, corresponded to about 0.130 lb./day of gain. Test the hypothesis that the profit derived from feeding B_{12} is zero. Ans $t = 1.84$, $P = 0.11$ (two-sided alternative)

4.6—Comparison of the means of two independent samples. When no pairing has been employed, we have two independent samples with means \bar{X}_1 , \bar{X}_2 , which are estimates of their respective population means μ_1 , μ_2 . Tests of significance and confidence intervals concerning the population difference, $\mu_1 - \mu_2$, are again based on the t -distribution, where t now has the value

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{\bar{X}_1 - \bar{X}_2}}$$

It is assumed that \bar{X}_1 and \bar{X}_2 are normally distributed and are independent. By theory, their difference is also normally distributed, so that the numerator of t is normal with mean zero.

The denominator of t is a sample estimate of the standard error of $(\bar{X}_1 - \bar{X}_2)$. The background for this estimate is given in the next two sections. First, we need an important new result for the population variance of a difference between any two variables X_1 and X_2 .

$$\sigma_{X_1 - X_2}^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2$$

The variance of a difference is the *sum* of the variances. This result holds for any two variables, whether normal or not, provided they are *independently* distributed.

4.7—The variance of a difference. A population variance is defined (section 2.12) as the average, over the population, of the squared deviations from the population mean. Thus we may write

$$\sigma_{X_1 - X_2}^2 = \text{Avg. of } \{(X_1 - X_2) - (\mu_1 - \mu_2)\}^2$$

But,

$$(X_1 - X_2) - (\mu_1 - \mu_2) = (X_1 - \mu_1) - (X_2 - \mu_2)$$

Hence, on squaring and expanding,

$$\begin{aligned} \{(X_1 - X_2) - (\mu_1 - \mu_2)\}^2 &= (X_1 - \mu_1)^2 + (X_2 - \mu_2)^2 \\ &\quad - 2(X_1 - \mu_1)(X_2 - \mu_2) \end{aligned}$$

Now average over all pairs of values X_1 , X_2 that can be drawn from their respective populations. By the definition of a population variance,

$$\begin{aligned}\text{Avg. of } (X_1 - \mu_1)^2 &= \sigma_{x_1}^2 \\ \text{Avg. of } (X_2 - \mu_2)^2 &= \sigma_{x_2}^2\end{aligned}$$

This leads to the general result

$$\sigma_{x_1 - x_2}^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 - 2 \text{ Avg. of } (X_1 - \mu_1)(X_2 - \mu_2) \quad (4.7.1)$$

At this point we use the fact that X_1 and X_2 are independently drawn. Because of this independence, any specific value of X_1 will appear with all the values of X_2 that can be drawn from its population. Hence, for this specific value of X_1 ,

$$\begin{aligned}\text{Avg. of } (X_1 - \mu_1)(X_2 - \mu_2) &= (X_1 - \mu_1)\{\text{Avg. of } (X_2 - \mu_2)\} \\ &= 0\end{aligned}$$

since μ_2 is the mean or average of all the values of X_2 . It follows that the overall average of the cross-product term $(X_1 - \mu_1)(X_2 - \mu_2)$ is zero, so that

$$\sigma_{x_1 - x_2}^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 \quad (4.7.2)$$

Apply this result to two means \bar{X}_1, \bar{X}_2 , drawn from populations with variance σ^2 . With samples of size n , each mean has variance σ^2/n . This gives

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = 2\sigma^2/n$$

The variance of a difference is twice the variance of an individual mean.

If σ is known, the preceding results provide the material for tests and confidence intervals concerning $\mu_1 - \mu_2$. To illustrate, from the table of pig gains (table 3.2.1) which we used to simulate a normal distribution with $\sigma = 10$ pounds, the first two samples drawn gave $\bar{X}_1 = 35.6$ and $\bar{X}_2 = 29.3$ pounds, with $n = 10$. Since the standard error of $\bar{X}_1 - \bar{X}_2$ is $\sqrt{2}\sigma/\sqrt{n}$, the quantity

$$Z = \sqrt{n}\{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)\}/\sqrt{2}\sigma$$

is a normal deviate. To test the null hypothesis that $\mu_1 = \mu_2$ we compute

$$Z = \frac{\sqrt{n}(\bar{X}_1 - \bar{X}_2)}{\sqrt{2}\sigma} = \frac{\sqrt{10}(6.3)}{\sqrt{2}(10)} = \frac{19.92}{14.14} = 1.41$$

From table A 3 a larger value of Z , ignoring sign, occurs about 16% of the trials. As we would expect, the difference is not significant. The 95% confidence limits for $(\mu_1 - \mu_2)$ are

$$(\bar{X}_1 - \bar{X}_2) \pm (1.96)\sqrt{2}\sigma/\sqrt{n}$$

4.8—A pooled estimate of variance. In most applications the value of σ^2 is not known. However, each sample furnishes an estimate of σ^2 : call these estimates s_1^2 and s_2^2 . With samples of the same size n , the best combined estimate is their pooled average $s^2 = (s_1^2 + s_2^2)/2$.

Since $s_1^2 = \Sigma x_1^2 / (n - 1)$ and $s_2^2 = \Sigma x_2^2 / (n - 1)$, where, as usual, $x_1 = X_1 - \bar{X}_1$ and $x_2 = X_2 - \bar{X}_2$, we may write

$$s^2 = \frac{\Sigma x_1^2 + \Sigma x_2^2}{2(n - 1)}$$

This formula is recommended for routine computing since it is quick, and extends easily to samples of unequal sizes.

The number of degrees of freedom in the pooled s^2 is $2(n - 1)$, the sum of the $d.f.$ in s_1^2 and s_2^2 . This leads to the result that

$$t = \sqrt{n} \{ (\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2) \} / \sqrt{2s}$$

follows Student's t -distribution with $2(n - 1)$ $d.f.$

The preceding analysis requires one additional assumption, namely that σ is the same in the two populations. The situations in which this assumption is suspect and the comparison of \bar{X}_1 and \bar{X}_2 when the assumption does not hold are discussed in section 4.14.

It is now time to apply these methods to a real experiment.

4.9—An experiment comparing two groups of equal size. Breneman (5) compared the 15-day mean comb weights of two lots of male chicks, one receiving sex hormone A (testosterone), the other C (dehydroandrosterone). Day-old chicks, 11 in number, were assigned at random to each of the treatments. To distinguish between the two lots, which were caged together, the heads of the chicks were stained red and purple respectively. The individual comb weights are recorded in table 4.9.1.

The calculations for the test of significance are given at the foot of the table. Note that in the Hormone A sample the correction term $(\Sigma X)^2/n$ is $(1,067)^2/11 = 103,499$. Note also the method recommended for computing the pooled s^2 . With 20 $d.f.$, the value of t is significant at the 1% level. Hormone A gives higher average comb weights than hormone C. The two sums of squares of deviations, 8,472 and 7,748, make the assumption of equal σ^2 appear reasonable.

The 95% confidence limits for $(\mu_1 - \mu_2)$ are

$$\bar{X}_1 - \bar{X}_2 \pm t_{0.05} s_{\bar{X}_1 - \bar{X}_2}$$

or, in this example,

$$41 - (2.086)(12.1) = 16 \text{ mg.}, \text{ and } 41 + (2.086)(12.1) = 66 \text{ mg.}$$

EXAMPLE 4.9.1—Lots of 10 bees were fed two concentrations of syrup, 20% and 65%, at a feeder half a mile from the hive (6). Upon arrival at the hive their honey sacs were removed and the concentration of the fluid measured. In every case there was a decrease from the feeder concentration. The decreases were: from the 20% syrup, 0.7, 0.5, 0.4, 0.7, 0.5, 0.4, 0.7, 0.4, 0.2, and 0.5; from the 65% syrup, 1.7, 2.8, 2.2, 1.4, 1.3, 2.1, 0.8, 3.4, 1.9, and 1.4%. Here, every observation in the second sample is larger than any in the first, so that rather obviously $\mu_1 < \mu_2$. Show that $t = 5.6$ if $\mu_1 - \mu_2 = 0$. There is little doubt

TABLE 4.9.1
TESTING THE DIFFERENCE BETWEEN THE MEANS OF TWO INDEPENDENT SAMPLES

Weight of Comb (mgs.)		
	Hormone A	Hormone C
	57	89
	120	30
	101	82
	137	50
	119	39
	117	22
	104	57
	73	32
	53	96
	68	31
	118	88
Totals	1,067	616
n	11	11
means	97	56
ΣX^2	111,971	42,244
$(\Sigma X)^2/n$	103,499	34,496
Σx^2	8,472	7,748
$d.f.$	10	10
Pooled s^2	$\frac{8,472 + 7,748}{10 + 10} = 811, \quad d.f. = 20$	
	$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{2s^2}{n}} = \sqrt{\frac{2(811)}{11}} = 12.14 \text{ mg.}$	
	$t = (\bar{X}_1 - \bar{X}_2)/s_{\bar{X}_1 - \bar{X}_2} = 41/12.14 = 3.38$	

that, under the experimental conditions imposed, the concentration during flight decreases more with the 65% syrup. But how about equality of variances? See sections 4.14 and 4.15 for further discussion.

EXAMPLE 4.9.2—Four determinations of the pH of Shelby loam were made with each of two types of glass electrode (7). With a modified quinhydrone electrode, the readings were 5.78, 5.74, 5.84, and 5.80; while with modified $Ag/AgCl$ electrode, they were 5.82, 5.87, 5.96, and 5.89. With the hypothesis that $\mu_1 - \mu_2 = 0$, calculate $t = 2.66$. Note: if you subtract 5.74 from every observation, the calculations are simpler.

EXAMPLE 4.9.3—In experiments to measure the effectiveness of carbon tetrachloride as a worm-killer, each of 10 rats received an injection of 500 larvae of the worm, *nippo-strongylus muris*. Eight days later 5 of the rats, chosen at random, each received 0.126 cc. of a solution of carbon tetrachloride, and two days later the rats were killed and the numbers of adult worms counted. These numbers were 378, 275, 412, 265, and 286 for the control rats and 123, 143, 192, 40, and 259 for the rats treated with CCl_4 . Find the significance probability for the difference in mean numbers of worms, and compute 95% confidence

limits for this difference. Ans. $t = 3.64$ with 8 *d.f.* P close to 0.01. Confidence limits are 63 and 280.

EXAMPLE 4.9.4—Fifteen kernels of mature Iodent corn were tested for crushing resistance. Measured in pounds the resistances were: 50, 36, 34, 45, 56, 42, 53, 25, 65, 33, 40, 42, 39, 43, 42. Another batch of 15 kernels was tested after being harvested in the dough stage: 43, 44, 51, 40, 29, 49, 39, 59, 43, 48, 67, 44, 46, 54, 64. Test the significance of the difference between the two means. Ans. $t = 1.38$.

EXAMPLE 4.9.5—In reading reports of researches it is sometimes desirable to supply a test of significance which was not considered necessary by the author. As an example, Smith (8) gave the sample mean yields and their standard errors for two crosses of maize as 8.84 ± 0.39 and 7.00 ± 0.18 grams. Each mean was the average of five replications. Determine if the mean difference is significant. Ans. $t = 4.29$, *d.f.* = 8. $P < 0.5\%$. To do this in the quickest way, satisfy yourself that the estimate of the variance of the difference between the two means is the sum of the squares of 0.39 and 0.18, namely 0.1845.

4.10—Groups of unequal sizes. Unequal numbers are common in comparisons made from survey data as, for example, comparing the mean incomes of men of similar ages who have master's and bachelor's degrees, or the severity of injury suffered in auto accidents by drivers wearing seat belts and drivers not wearing seat belts. In planned experiments, equal numbers are preferable, being simpler to analyze and more efficient, but equality is sometimes impossible or inconvenient to attain. Two lots of chicks from two batches of eggs treated differently nearly always differ in the number of birds hatched. Occasionally, when a new treatment is in short supply, an experiment with unequal numbers is set up deliberately.

Unequal numbers occur also in experiments because of accidents and losses during the course of the trial. In such cases the investigator should always consider whether any loss represents a failure of the treatment rather than an accident that is not to be blamed on the treatment. Needless to say, such situations require careful judgment.

The statistical analysis for groups of unequal sizes follows almost exactly the same pattern as that for groups of equal sizes. As before, we assume that the variance is the same in both populations unless otherwise indicated. With samples of sizes n_1 , n_2 , their means \bar{X}_1 and \bar{X}_2 have variances σ^2/n_1 and σ^2/n_2 . The variance of the difference is then

$$\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2(1/n_1 + 1/n_2) = \sigma^2 \left(\frac{n_1 + n_2}{n_1 n_2} \right)$$

In order to form a pooled estimate of σ^2 , we follow the rule given for equal-sized samples. *Add the sums of squares of deviations in the numerators of s_1^2 and s_2^2 , and divide by the sum of their degrees of freedom.* These degrees of freedom are $(n_1 - 1)$ and $(n_2 - 1)$, so that the denominator of the pooled s^2 is $(n_1 + n_2 - 2)$. This quantity is also the number of *d.f.* in the pooled s^2 . The procedure will be clear from the example in table 4.10.1. Note how closely the calculations follow those given in table 4.9.1 for samples of equal sizes.

TABLE 4.10.1
ANALYSIS FOR TWO SAMPLES OF UNEQUAL SIZES. GAINS IN WEIGHTS OF TWO LOTS
OF FEMALE RATS (28-84 days old) UNDER TWO DIETS

	Gains (gms.)	
	High Protein	Low Protein
	134	70
	146	118
	104	101
	119	85
	124	107
	161	132
	107	94
	83	
	113	
	129	
	97	
	123	
Totals	1440	707
n	12	7
means	120	101
ΣX^2	177,832	73,959
$(\Sigma X)^2/n$	172,800	71,407
Σx^2	5,032	2,552
df	11	6
Pooled s^2	$\frac{5,032 + 2,552}{11 + 6} = 446.12, \quad df = 17$	

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 \left(\frac{n_1 + n_2}{n_1 n_2} \right)} = \sqrt{\{ (446.12)(19)/84 \}} = 10.04 \text{ gms.}$$

$$t = 19/10.04 = 1.89, P \text{ about } 0.08$$

The high protein diet showed a slightly greater mean gain. Since P is about 0.08, however, a difference as large as the observed one would occur about 1 in 12 times by chance, so that the observed difference cannot be regarded as established by the usual standards in tests of significance.

For evidence about homogeneity of variance in the two populations, observe that $s_1^2 = 5,032/11 = 457$ and $s_2^2 = 2,552/6 = 425$.

If the investigator is more interested in estimates than in tests, he may prefer the confidence interval. He reports an observed difference of 19 gms. in favor of the high protein diet, with 95% confidence limits -2.2 and 40.2 gms.

EXAMPLE 4.10.1—The following are the rates of diffusion of carbon dioxide through two soils of different porosity (9). Through a fine soil (f) 20, 31, 18, 23, 23, 28, 23, 26, 27, 24, 42, 17, 25; through a coarse soil (c) 19, 30, 32, 28, 15, 26, 35, 18, 25, 27, 35, 34. Show

that pooled $s^2 = 35.83$, $s_{\bar{y}_1 - \bar{y}_2} = 2.40$, $d.f. = 23$, and $t = 1.67$. The difference, therefore, is not significant.

EXAMPLE 4.10.2—The total nitrogen content of the blood plasma of normal albino rats was measured at 37 and 180 days of age (10). The results are expressed as gms. per 100 cc. of plasma. At age 37 days, 9 rats had 0.98, 0.83, 0.99, 0.86, 0.90, 0.81, 0.94, 0.92, and 0.87, at age 180 days, 8 rats had 1.20, 1.18, 1.33, 1.21, 1.20, 1.07, 1.13, and 1.12 gms. per 100 cc. Since significance is obvious, set a 95% confidence interval on the population mean difference. Ans. 0.21 to 0.35 gms./100 cc.

EXAMPLE 4.10.3—Sometimes, especially in comparisons made from surveys, the two samples are large. Time is saved by forming frequency distributions and computing the means and variances as in section 3.11. The following data from an experiment serve as an illustration. The objective was to compare the effectiveness of two antibiotics, A and B, for treating patients with lobar pneumonia. The numbers of patients were 59 and 43. The data are the numbers of days needed to bring the patient's temperature down to normal.

No. of Days		1	2	3	4	5	6	7	8	9	10	Total
No. of Patients	A	17	8	5	9	7	1	2	1	2	7	59
	B	15	8	8	5	3	1	0	0	0	3	43

What are your conclusions about the relative effectiveness of the two antibiotics in bringing down the fever? Ans. The difference of about 1 day in favor of B has a P value between 0.05 and 0.025. Note that although these are frequency distributions, the only real grouping is in the 10-day groups, which actually represented "at least 10" and were arbitrarily rounded to 10. Since the distributions are very skew, the analysis leans heavily on the Central Limit Theorem. Do the variances given by the two drugs appear to differ?

EXAMPLE 4.10.4—Show that if the two samples are of sizes 6 and 12, the S.D. of the difference in means is the same as when the samples are both of size 8. Are the $d.f.$ in the pooled s^2 the same?

EXAMPLE 4.10.5—Show that the pooled s^2 is a weighted mean of s_1^2 and s_2^2 in which each is weighted by its number of $d.f.$

4.11—Paired versus independent groups. The formula for the variance of a difference throws more light on the circumstances in which pairing is effective. Quoting formula (4.7.1),

$$\sigma_{X_1 - X_2}^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 - 2 \text{ Avg. of } (X_1 - \mu_1)(X_2 - \mu_2)$$

When pairing, we try to choose pairs such that if X_1 is high, so is X_2 . Thus, if $(X_1 - \mu_1)$ is positive, so is $(X_2 - \mu_2)$, and their product $(X_1 - \mu_1)(X_2 - \mu_2)$ is positive. Similarly, in successful pairing, when $(X_1 - \mu_1)$ is negative, $(X_2 - \mu_2)$ will usually also be negative. Their product $(X_1 - \mu_1)(X_2 - \mu_2)$ is again positive. For paired samples, then, the average of this product is positive. This helps, because it makes the variance of $(X_1 - X_2)$ less than the sum of their variances, sometimes very much less. The average value of the product over the population is called the *covariance* of X_1 and X_2 , and is studied in chapter 7. The result for the variance of a difference may now be written

$$\sigma_{X_1 - X_2}^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 - 2 \text{ Cov. } (X_1, X_2)$$

Pairing is not always effective, because X_1 and X_2 may be poorly correlated. Fortunately, it is possible from the results of a paired experiment to estimate what the standard error of $(\bar{X}_1 - \bar{X}_2)$ would have been if the experiment had been conducted as two independent groups. By this calculation the investigator can appraise the success of his pairing, which guides him in deciding whether the pairing is worth continuing in future experiments.

With paired samples of size n , the standard error of the mean difference $\bar{D} = \bar{X}_1 - \bar{X}_2$ is σ_D/\sqrt{n} , where σ_D is the standard deviation of the population of paired differences (section 4.3). For an experiment with two independent groups, the standard error of $\bar{X}_1 - \bar{X}_2$ is $\sqrt{2}\sigma/\sqrt{n}$, where σ is the standard deviation of the original population from which we drew the sample of size $2n$ (section 4.7). Omitting the \sqrt{n} , the quantities that we want to compare are σ_D and $\sqrt{2}\sigma$. Usually, the comparison is made in terms of variances: we compare σ_D^2 with $2\sigma^2$.

From the statistical analysis of the paired experiment, we have an unbiased estimate s_D^2 of σ_D^2 . The problem is to obtain an estimate of $2\sigma^2$. One possibility is to analyze the results of the paired experiment by the method of section 4.9 for two independent samples, using the pooled s^2 as an estimate of σ^2 . This procedure gives a good approximation when n is large, but is slightly wrong, because the two samples from which s^2 was computed were not independent. An unbiased estimate of $2\sigma^2$ is given by the formula

$$2\hat{\sigma}^2 = 2s^2 - (2s^2 - s_D^2)/(2n - 1)$$

(The 'hat' [$\hat{}$] placed above a population parameter is often used in mathematical statistics to denote an estimate of that parameter.)

Let us apply this method to the paired experiment on virus lesions (table 4.3.1, p. 95), which gave $s_D^2 = 18.57$. You may verify that the pooled s^2 is 45.714, giving $2s^2 = 91.43$. Hence, an unbiased estimate of $2\sigma^2$ is

$$2\hat{\sigma}^2 = 91.43 - (91.43 - 18.57)/15 = 86.57$$

The pairing has given a much smaller variance of the mean difference, $18.57/n$ versus $86.57/n$. What does this imply in practical terms? With independent samples, the sample size would have to be increased from 8 pairs to $8(86.57)/(18.57)$, or about 37 pairs, in order to give the same variance of the mean difference as does the paired experiment. The saving in amount of work due to pairing is large in this case.

The computation overlooks one point. In the paired experiment, s_D^2 has 7 *d.f.*, whereas the pooled s^2 would have 14 *d.f.* for error. The *t*-value used in tests of significance or in computing confidence limits would be slightly smaller with independent samples than with paired samples. Several writers (11), (12), (13), have discussed the allowance that should be made for this difference in number of *d.f.* We suggest a

rule given by Fisher (12). Multiply the estimated variance by $(f + 3)/(f + 1)$, where f is the *d.f.* that the experimental plan provides. Thus we compare

$$(18.57)(10)/8 = 23.2, \text{ with } (86.57)(17)/(15) = 98.1$$

D. R. Cox (13) suggests the multiplier $(f + 1)^2/f^2$. This gives almost the same results, imposing a slightly higher penalty when f is small.

From a single experiment a comparison like the above is not very precise, particularly if n is small. The results of several paired experiments in which the same criterion for pairing was employed give a more accurate picture of the success of the pairing. If the criterion has no correlation with the response variable, there is a small loss in accuracy from pairing due to the adjustment for *d.f.* There may even be a substantial loss in accuracy if the criterion is badly chosen so that members of a pair are negatively correlated.

When analyzing the results of a comparison of two procedures, the investigator must know whether his samples are paired or independent and must use the appropriate analysis. Sometimes a worker with paired data forgets this when it comes to analysis, and carries out the statistical analysis as if the two samples were independent. This is a serious mistake if the pairing has been effective. In the virus lesions example, he would be using $2s^2/n$ or $91.43/8 = 11.44$ as the variance of \bar{D} instead of $18.57/8 = 2.32$. The mistake throws away all the advantage of the pairing. Differences that are actually significant may be found non-significant, and confidence intervals will be too wide.

Analysis of independent samples as if they were paired seems to be rare in practice. If the members of each sample are in essentially random order, so that the pairs are a random selection, the computed s_p^2 may be shown to be an unbiased estimate of $2\sigma^2$. Thus the analysis still provides an unbiased estimate of the variance of $(\bar{X}_1 - \bar{X}_2)$ and a valid *t*-test. There is a slight loss in sensitivity, since *t*-tests are based on $(n - 1)$ *d.f.*, instead of $2(n - 1)$ *d.f.*

As regards assumptions, pairing has the advantage that its *t*-test does not require $\sigma_1 = \sigma_2$. "Random" pairing of independent samples has been suggested as a means of obtaining tests and confidence limits when the investigator knows that σ_1 and σ_2 are unequal.

Artificial pairing of the results, by arranging each sample in descending order and pairing the top two, the next two, and so on, produces a great under-estimation of the true variance of \bar{D} . This effect may be illustrated by the first two random samples of pig gains from table 3.3.1 (p. 69). The population variance σ^2 is 100, giving $2\sigma^2 = 200$. In table 4.11.1 this method of artificial pairing has been employed.

Instead of the correct value of 200 for $2\sigma^2$ we get an estimate s_D^2 of only 8.0. Since $s_{\bar{D}} = \sqrt{(8.0/10)} = 0.894$, the *t*-value for testing \bar{D} is $t = 6.3/0.894 = 7.04$, with 9 *d.f.* This gives a *P* value of much less than 0.1%, although the two samples were drawn from the same population.

TABLE 4.11 1
TWO SAMPLES OF 10 PIG GAINS ARRANGED IN DESCENDING ORDER, TO ILLUSTRATE
THE ERRONEOUS CONCLUSIONS FROM ARTIFICIAL PAIRING

Sample 1	57	53	39	39	36	34	33	29	24	12	Mean = 35.6
Sample 2	53	44	32	31	30	30	24	19	19	11	Mean = 29.3
Diff.	4	9	7	8	6	4	9	10	5	1	Mean = 6.3
$\Sigma d^2 = 469 - (63)^2/10 = 72.1, s_p^2 = 72.1/9 = 8.0$											

EXAMPLE 4.11.1—In planning experiments to test the effects of two pain-killers on the ability of young men to tolerate pain from a narrow beam of light directed at the arm, each subject was first rated several times as to the amount of heat energy that he bore without complaining of discomfort. The subjects were then paired according to these initial scores. In a later experiment the amounts of energy received at the point at which the subject complained were as follows, A and B denoting the treatments.

Pair	1	2	3	4	5	6	7	8	9	Sums
A	15	2	4	1	5	7	1	0	-3	32
B	6	7	3	4	3	2	3	0	-6	22

To simplify calculations, 30 was subtracted from each original score. Show that for appraising the effectiveness of the pairing, comparable variances are 22.5 for the paired experiment and 44.6 for independent groups (after allowing for the difference in *d.f.*). The preliminary work in rating the subjects reduced the number of subjects needed by almost one-half.

EXAMPLE 4.11.2—In a previous experiment comparing two routes A and B for driving home from an office (example 4.3.4), pairing was by days of the week. The times taken (- 23 mins.) for the ten pairs were as follows:

A	5.7	3.2	1.8	2.3	2.1	0.9	3.1	2.8	7.3	8.4
B	2.4	2.8	1.9	2.0	0.9	0.3	3.6	1.8	5.8	7.3
Diff.	3.3	0.4	-0.1	0.3	1.2	0.6	-0.5	1.0	1.5	1.1

Show that if the ten nights on which route A was used had been drawn at random from the twenty nights available, the variance of the mean difference would have been about 8 times as high as with this pairing

EXAMPLE 4.11.3 If pairing has not reduced the variance, so that $s_p^2 = 2\sigma^2$, show that allowance for the error *d.f.* by Fisher's rule makes pairing 15% less effective than independent groups when $n = 5$ and 9% less effective when $n = 10$. In small experiments, pairing is inadvisable unless a sizeable reduction in variance is expected.

4.12—Precautions against bias-randomization. With either independent or paired samples, the analysis assumes that the difference $(\bar{X}_1 - \bar{X}_2)$ is an unbiased estimate of the population mean difference between the two treatments. Unless precautions are taken when conducting an experiment, $(\bar{X}_1 - \bar{X}_2)$ may be subject to a bias of unknown

amount that makes the conclusion false. Corner (14) describes an example in which, when picking rabbits out of a hatch, one worker tended to pick large rabbits, another to pick small rabbits, although neither was aware of his personal bias. If the rabbits for treatment A are picked out first, a bias will be introduced if the final response depends on the weight of the rabbit. If the animals receiving treatment A are kept in one cage and those having B in another, temperature, draftiness, or sources of infection in one cage may affect all the animals receiving A differently from those receiving B. When the application of the treatment or the measurement of response takes considerable time, unsuspected time trends may be present, producing bias if all replicates of treatment A are processed first. The investigator must be constantly on guard against such sources of bias.

One helpful device, now commonly used, is *randomization*. When pairs have been formed, the decision as to which member of a pair receives treatment A is made by tossing a coin or by using a table of random numbers. If the random number drawn is odd, the first member of the pair will receive treatment A. With 10 pairs, we draw 10 random digits from table A 1, say 9, 8, 0, 1, 8, 3, 6, 8, 0, 3. In pairs 1, 4, 6, and 10, treatment A is given to the first member of the pair and B to the second member. In the remaining pairs, the first member receives B.

With independent samples, random numbers are used to divide the $2n$ subjects into two groups of n . Number the subjects in any order from 1 to $2n$. Proceed down a column of random numbers, allotting the subject to A if the number is odd, to B if even, continuing until n A's or n B's have been allotted. With 14 subjects and the same random numbers as above, subjects 1, 4, 6, and 10 receive A and subjects 2, 3, 5, 7, 8, and 9 receive B. Thus far we have allotted four A's and six B's, so that more random numbers must be drawn. The next two in the column are 1, 8. Subject 11 gets A and subject 12 gets B. Since seven B's have been assigned we stop, giving A to subjects 13 and 14.

Randomization gives each treatment an equal chance of being allotted to any subject that happens to give an unusually good or unusually poor response, exactly as assumed in the theory of probability on which the statistical analysis is based. Randomization does not guarantee to balance out the natural differences between the members of a pair exactly. With n pairs, there is a small probability, $1/2^{n-1}$, that one treatment will be assigned to the superior member in every pair. With 10 pairs this probability is about 0.002. If the experimenter can predict which is likely to be the superior member in each pair, he should try a more sophisticated design (chapter 11) that utilizes this information more effectively than randomization. Randomization serves primarily to protect against sources of bias that are *unsuspected*. Randomization can be used not merely in the allocation of treatments to subjects, but at any later stage in which it may be a safeguard against bias, as discussed in (11), (13).

Both independent and paired samples are much used in comparisons

made from surveys. The problem of avoiding misleading conclusions is formidable with survey data (15). Suppose we tried to learn something about the value of completing a high school education by comparing, some years later, the incomes, job satisfaction, and general well-being of a group of boys who completed high school with a group from the same schools who started but did not finish. Obviously, significant differences found between the sample means may be due to factors other than the completion of high school in itself: differences in the natural ability and personal characteristics of the boys, in the parents' economic level and number of useful contacts, and so on. Pairing the subjects on their school performance and parents' economic level helps, but no randomization within pairs is possible, and a significant mean difference may still be due to extraneous factors whose influence has been overlooked.

Remember that a significant t -value is evidence that the population *means* differ. Popular accounts are sometimes written as if a significant t implies that every member of population 1 is superior to every member of population 2. "The oldest child in the family achieves more in science or in business." In fact, the two populations may largely overlap even though t is significant.

4.13—Sample size in comparative experiments. In planning an experiment to compare two treatments, the following method is often used to estimate the size of sample needed. The investigator first decides on a value δ which represents the size of difference between the true effect of the treatments that he regards as important. If the true difference is as large as δ , he would like the experiment to have a high probability of showing a statistically significant difference between the treatment means. Probabilities of 0.80 and 0.90 are common. A higher probability, say 0.95 or 0.99, can be set, but the sample size required to meet these severer specifications is often too expensive.

This way of stating the aims in planning the sample size is particularly appropriate when (i) the treatments are a standard treatment and a new treatment that the experimenter hopes will be better than the standard, and (ii) he intends to discard the new treatment if the experiment does not show it to be significantly superior to the standard. In these circumstances he does not mind dropping the new treatment if it is at most only slightly better than the standard, but he does not want to drop it, on the evidence of the experiment, if it is substantially superior. The value of δ measures his idea of a substantial true difference.

In order to make the calculation the experimenter supplies:

1. the value of δ ,
2. the desired probability P' of obtaining a significant result if the true difference is δ ,
3. the significance level α of the test, which may be either one-tailed or two-tailed.

Consider paired samples. Assume at first that σ_D is known and that

the test is two-tailed. In our specification, the observed mean difference $\bar{D} = \bar{X}_1 - \bar{X}_2$ is normally distributed about δ with standard deviation σ_D/\sqrt{n} . This distribution is shown in figure 4.13.1, which forms the basis of our explanation. We have assumed $\delta > 0$.

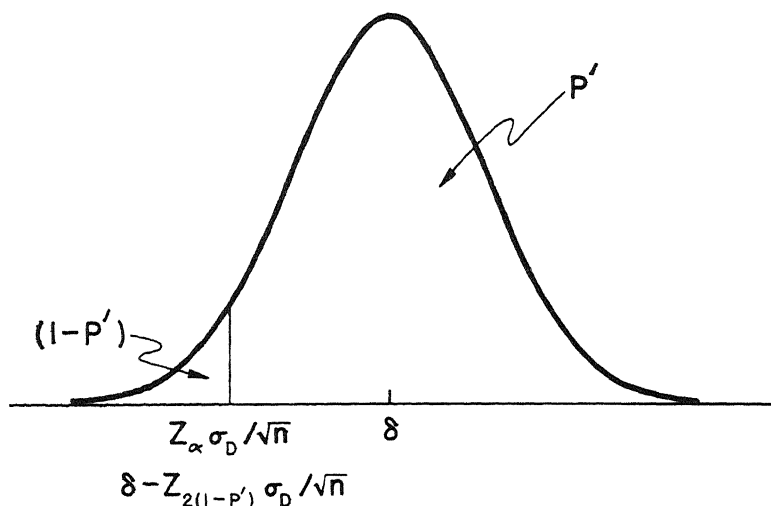


FIG. 4.13.1—Frequency distribution of the mean difference \bar{D} between two treatments.

In order to be statistically significant, \bar{D} must exceed $Z_{\alpha} \sigma_D / \sqrt{n}$, where Z_{α} is the normal deviate corresponding to the two-tailed significance level α . (For $\alpha = 0.01, 0.05, 0.10$, the values of Z_{α} are 2.576, 1.960, and 1.645, respectively.) The vertical line in figure 4.13.1 shows the critical value.

In our specification, the probability that \bar{D} exceeds this value must be P' . That is, this value divides the frequency distribution of \bar{D} into an area P' on the right and $(1 - P')$ on the left. Consider the standard normal curve, with mean 0 and S.D. 1. With $P' > 1/2$, the point at which the area on the left is $(1 - P')$ is *minus* the normal deviate corresponding to a *one-tailed* significance level $(1 - P')$. This is the same as *minus* the normal deviate corresponding to a *two-tailed* significance level $2(1 - P')$, or in our notation to $-Z_{2(1-P')}$. For instance, with $P' = 0.9$, this is the normal deviate $-Z_{0.2}$, or -1.282 .

Since \bar{D} has mean δ and S.D. σ_D/\sqrt{n} , the quantity $(\bar{D} - \delta) (\sigma_D/\sqrt{n})$ follows the standard normal curve. Hence, the value of \bar{D} that is exceeded with probability P' is given by the equation

$$\frac{\bar{D} - \delta}{\sigma_D / \sqrt{n}} = -Z_{2(1-P')}$$

or,

$$\bar{D} = \delta - Z_{2(1-P')} \sigma_D / \sqrt{n}$$

It follows that our specification is satisfied if

$$Z_\alpha \sigma_D / \sqrt{n} = \delta - Z_{2(1-P')} \sigma_D / \sqrt{n}$$

A look at figure 4.13.1 may help at this point. Write $\beta = 2(1 - P')$ and solve for n .

$$n = (Z_\alpha + Z_\beta)^2 \sigma_D^2 / \delta^2 \quad (4.13.1)$$

To illustrate, for a one-tailed test at the 5% level with $P' = 0.90$, we have $Z_\alpha = 1.645$, $Z_\beta = 1.282$, giving $n = 8.6 \sigma_D^2 / \delta^2$. Note that n is the size of each sample, the total number of observations being $2n$.

Formula (4.13.1) for n remains the same for independent samples, except that σ_D^2 is replaced by $2\sigma^2$.

The two-tailed case involves a slight approximation. In a two-tailed test, \bar{D} in figure 4.13.1 is also significant if it is less than $-Z_{\alpha/2} \sigma_D / \sqrt{n}$. But with δ positive, the probability that this happens is negligible in most practical situations.

Table 4.13.1 presents the multipliers $(Z_\alpha + Z_\beta)^2$ that are most frequently used.

When σ_D and σ are estimated from the results of the experiment, t -tests replace the normal deviate tests. The logical basis of the argument remains the same, but the formula for n becomes an integral equation in calculus that must be solved by successive approximation. This equation was given by Neyman (21) to whom this method of determining sample size is due.

For practical purposes, the following approximation agrees well enough with the values of n as found from Neyman's solution:

1. Find n_1 to one decimal place by table 4.13.1.

TABLE 4.13.1
MULTIPLIERS OF σ_D^2 / δ^2 IN PAIRED SAMPLES, AND OF $2\sigma^2 / \delta^2$ IN INDEPENDENT SAMPLES,
IN ORDER TO DETERMINE THE SIZE OF EACH SAMPLE

P'	Two-tailed Tests			One-tailed Tests		
	0.01	Level 0.05	0.10	0.01	Level 0.05	0.10
0.80	11.7	7.9	6.2	10.0	6.2	4.5
0.90	14.9	10.5	8.6	13.0	8.6	6.6
0.95	17.8	13.0	10.8	15.8	10.8	8.6

2. Calculate f , the number of degrees of freedom supplied by an experiment of this size (rounding n_1 upwards for this step).

3. Multiply n_1 in step 1 by $(f + 3)/(f + 1)$.

To illustrate, suppose that a 10% difference δ is regarded as important and that $P' = 0.80$ in a two-tailed 5% test of significance. The samples are to be independent, and past experience has shown that σ is about 6%. The multiplier for $P' = 0.80$ and a 5% two-tailed test in table 4.13.1 is 7.9. Since $2\sigma^2/\delta^2 = 72/100 = 0.72$, $n_1 = (7.9)(0.72) = 5.7$. With a sample size of 6 in each group, $f = 10$. Hence we take $n = (13)(5.7)/11 = 6.7$, which we round up to 7.

Note that the experimenter must still guess a value of σ_D or σ . Usually it is easier to guess σ . If pairing is to be used but is expected to be only moderately effective, take $\sigma_D = \sqrt{2} \sigma$, reducing this value if something more definite is known about the effectiveness of pairing. This uncertainty is the chief source of inaccuracy in the process.

The preceding method is designed to protect the investigator against finding a non-significant result and consequently dropping a new treatment that is actually effective, because his experiment was too small. The method is therefore most useful in the early stages of a line of work. At later stages, when something has been learned about the sizes of differences produced by new treatments, we may wish to specify the size of the standard error or the half-width of the confidence interval that will be attached to an estimated difference.

For example, previous small experiments have indicated that a new treatment gives an increase of around 20%, and σ is around 7%. The investigator would like to estimate this increase, in his next experiment, with a standard error of $\pm 2\%$. He sets $\sqrt{2(7)}/\sqrt{n} = 2$, giving $n = 25$ in each group. This type of rough calculation is often helpful in later work.

EXAMPLE 4 13.1—In table 4.13 1, verify the multipliers given for a one-tailed test at the 1% level with $P' = 0.90$ and for a two-tailed test at the 10% level with $P' = 0.80$

EXAMPLE 4 13.2—In planning a paired experiment, the investigator proposes to use a one-tailed test of significance at the 5% level, and wants the probability of finding a significant difference to be 0.90 if (i) $\delta = 10\%$, (ii) $\delta = 5\%$. How many pairs does he need? In each case, give the answer if (a) σ_D is known to be 12% , (b) σ_D is guessed as 12% , but a t -test will be used in the experiment. Ans (ia) 13, (ib) 15, (iia) 50, (iib) 52

EXAMPLE 4 13.3—In the previous example, how many pairs would you guess to be necessary if $\delta = 2.5\%$? The answer brings out the difficulty of detecting small differences in comparative experiments with variable data

EXAMPLE 4 13.4—If $\sigma_D = 5$, how many pairs are needed to make the half-width of the 90% confidence interval for the difference between the two population means $= 2$? Ans $n = 17$

4.14—Analysis of independent samples when $\sigma_1 \neq \sigma_2$. The ordinary method of finding confidence limits and making tests of significance for the difference between the means of two independent samples assumes that

the two population variances are the same. Common situations in which the assumption is suspect are as follows:

(1) When the samples come from populations of different types, as in comparisons made from survey data. In comparing the average values of some characteristic of boys from public and private schools, we might expect, from our knowledge of the differences in the two kinds of schools, that the variances will not be the same.

(2) When computing confidence limits in cases in which the population means are obviously widely different. The frequently found result that σ tends to change, although slowly, when μ changes, would make us hesitant to assume $\sigma_1 = \sigma_2$.

(3) With samples from populations that are markedly skew. In many such populations the relation between σ and μ is often relatively strong.

When $\sigma_1 \neq \sigma_2$, the formula for the variance of $(\bar{X}_1 - \bar{X}_2)$ in independent samples still holds, namely,

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

The two samples furnish unbiased estimates s_1^2 of σ_1^2 and s_2^2 of σ_2^2 . Consequently, the ordinary t is replaced by the quantity

$$t' = (\bar{X}_1 - \bar{X}_2) \sqrt{(s_1^2/n_1 + s_2^2/n_2)}$$

This quantity does not follow Student's t -distribution when $\mu_1 = \mu_2$. Two different forms of the distribution of t' , arising from different theoretical backgrounds, have been worked out, one due to Behrens (16) and Fisher (17), the other to Welch and Aspin (18), (22). Both require special tables, given in the references. The tables differ relatively little, the Behrens-Fisher table being on the whole more conservative, in the sense that slightly higher values of t' are required for significance. The following approximation due to Cochran (19), which uses the ordinary t -table, is sufficiently accurate for our purposes. It is usually slightly more conservative than the Behrens-Fisher solution.

Case 1: $n_1 = n_2$. With $n_1 = n_2 = n$, the variance in the denominator of t' is $(s_1^2 + s_2^2)/n$. But this is just $2s^2/n$, where s^2 is the pooled variance. Thus, in this case, $t' = t$. The rule is: calculate t in the usual way, but give it $(n - 1)$ d.f. instead of $2(n - 1)$.

Case 2: $n_1 \neq n_2$. Calculate t . To find its significance level, look up the significance levels of t in table A 4 for $(n_1 - 1)$ and $(n_2 - 1)$ d.f. Call these values t_1 and t_2 . The significance level of t' is, approximately,

$$(w_1 t_1 + w_2 t_2)/(w_1 + w_2), \text{ where } w_1 = s_1^2 n_1, w_2 = s_2^2 n_2$$

The following artificial examples illustrates the calculations: A quick but imprecise method of estimating the concentration of a chemical in a vat has been developed. Eight samples from the vat are analyzed, as well

as four samples by the standard method, which is precise but slow. In comparing the means we are examining whether the quick method gives a systematic over- or underestimate. Table 4.14.1 gives the computations.

TABLE 4.14.1
A TEST OF $(\bar{X}_1 - \bar{X}_2)$ WHEN $\sigma_1 \neq \sigma_2$
CONCENTRATION OF A CHEMICAL BY TWO METHODS

Standard		Quick
	25	23
	24	18
	25	22
	26	28
		17
		25
		19
		16
$\bar{X}_1 = 25$		$\bar{X}_2 = 21$
$n_1 = 4$		$n_2 = 8$
$s_1^2 = 0.67$		$s_2^2 = 17.71$
$s_1^2/n_1 = 0.17$		$s_2^2/n_2 = 2.21$
$t' = 4/\sqrt{2.38} = 2.60$		
$t_1(3 \text{ d.f.}) = 3.182$		$t_2(7 \text{ d.f.}) = 2.365$
$t_{0.05} = 5\% \text{ level of } t' = \{(0.17)(3.182) + (2.21)(2.365)\}/2.38$		
$= 2.42$		

Since $2.60 > 2.42$, the difference is significant at the 5% level; the quick method appears to underestimate.

Approximate 95% confidence limits for $(\mu_1 - \mu_2)$ are

$$\bar{X}_1 - \bar{X}_2 \pm t'_{0.05} s_{\bar{X}_1 - \bar{X}_2}$$

or in this example, $4 \pm (2.42)(1.54) = 4 \pm 3.7$.

The ordinary t -test with a pooled s^2 gives $t = 1.84$, to which we would erroneously attribute 10 d.f. The t -test tends to give too few significant results when the larger sample has the larger variance, as in this example, and too many when the larger sample has the smaller variance.

Sometimes, when it seemed reasonable to assume that $\sigma_1 = \sigma_2$ or when the investigator failed to think about the question in advance, he notices that s_1^2 and s_2^2 are distinctly different. A test of the null hypothesis that $\sigma_1 = \sigma_2$, given in the next section, is useful. If the null hypothesis is rejected, the origin of the data should be re-examined. This may reveal some cause for expecting the standard deviations to be different. In case of doubt it is better to avoid the assumption that $\sigma_1 = \sigma_2$.

4.15—A test of the equality of two variances. The null hypothesis is that s_1^2 and s_2^2 are independent random samples from normal populations with the same variance σ^2 . In situations in which there is no prior

reason to anticipate inequality of variance, the alternative is a two-sided one: $\sigma_1 \neq \sigma_2$. The test criterion is $F = s_1^2/s_2^2$, where s_1^2 is the larger mean square. The distribution of F when the null hypothesis is true was worked out by Fisher (20) early in the 1920's. Like χ^2 and t it is one of the basic distributions in modern statistical methods. A condensed two-tailed table of the 5% significance levels of F is table 4.15.1.

TABLE 4.15.1
5% LEVEL (TWO-TAILED) OF THE DISTRIBUTION OF F

$f_2 = d.f.$ for Smaller Mean Square	$f_1 = d.f.$ for Larger Mean Square									
	2	4	6	8	10	12	15	20	30	∞
2	39.00	39.25	39.33	39.37	39.40	39.42	39.43	39.45	39.46	39.50
3	16.04	15.10	14.74	14.54	14.42	14.34	14.25	14.17	14.08	13.90
4	10.65	9.60	9.20	8.98	8.84	8.75	8.66	8.56	8.46	8.26
5	8.43	7.39	6.98	6.76	6.62	6.52	6.43	6.33	6.23	6.02
6	7.26	6.23	5.82	5.60	5.46	5.37	5.27	5.17	5.07	4.85
7	6.54	5.52	5.12	4.90	4.76	4.67	4.57	4.47	4.36	4.14
8	6.06	5.05	4.65	4.43	4.30	4.20	4.10	4.00	3.89	3.67
9	5.71	4.72	4.32	4.10	3.96	3.87	3.77	3.67	3.56	3.33
10	5.46	4.47	4.07	3.85	3.72	3.62	3.52	3.42	3.31	3.08
12	5.10	4.12	3.73	3.51	3.37	3.28	3.18	3.07	2.96	2.72
15	4.76	3.80	3.41	3.20	3.06	2.96	2.86	2.76	2.64	2.40
20	4.46	3.51	3.13	2.91	2.77	2.68	2.57	2.46	2.35	2.09
30	4.18	3.25	2.87	2.65	2.51	2.41	2.31	2.20	2.07	1.79
∞	3.69	2.79	2.41	2.19	2.05	1.94	1.83	1.71	1.57	1.00

Use of the table is illustrated by the bee data in example 4.9.1. Bees fed a 65% concentration of syrup showed a mean decrease in concentration of 1.9%, with $s_1^2 = 0.589$, while bees fed a 20% concentration gave a mean decrease of 0.5% with $s_2^2 = 0.027$. Each mean square has 9 $d.f.$ Hence

$$F = 0.589/0.027 = 22.1$$

In the row for 9 $d.f.$ and the column for 9 $d.f.$ (interpolated between 8 and 10) the 5% level of F is 4.03. The null hypothesis is rejected. No clear explanation of the discrepancy in variances was found, except that it may reflect the association of a smaller variance with a smaller mean. The difference between the means is strongly significant whether the variances are assumed the same or not.

Often a one-tailed test is wanted, because we know, in advance of seeing the data, which population will have the higher variance if the null hypothesis is untrue. The numerator of F is s_1^2 if $\sigma_1 > \sigma_2$ is the alternative, and s_2^2 if $\sigma_2 > \sigma_1$ is the alternative. Table A 14 presents one-tailed levels of F directly.

EXAMPLE 4.15.1—Young examined the basal metabolism of 26 college women in two groups of $n_1 = 15$ and $n_2 = 11$; $\bar{X}_1 = 34.45$ and $\bar{X}_2 = 33.57$ cal./sq. m./hr.; $\Sigma x_1^2 = 69.36$, $\Sigma x_2^2 = 13.66$. Test $H_0: \sigma_1 = \sigma_2$. Ans. $F = 3.62$ to be compared with $F_{0.05} = 3.55$. (Data from Ph.D. thesis, Iowa State University, 1940).

BASAL METABOLISM OF 26 COLLEGE WOMEN
(Calories per square meter per hour)

7 or More Hours of Sleep				6 or Less Hours of Sleep			
1.	35.3	9.	33.3	1.	32.5	7.	34.6
2.	35.9	10.	33.6	2.	34.0	8.	33.5
3.	37.2	11.	37.9	3.	34.4	9.	33.6
4.	33.0	12.	35.6	4.	31.8	10.	31.5
5.	31.9	13.	29.0	5.	35.0	11.	33.8
6.	33.7	14.	33.7	6.	34.6		
7.	36.0	15.	35.7				$\Sigma X_2 = 369.3$
8.	35.0		$\Sigma X_1 = 516.8$				
$\bar{X}_1 = 34.45$ cal./sq. m./hr.				$\bar{X}_2 = 33.57$ cal./sq. m./hr			

EXAMPLE 4.15.2—In the metabolism data there is little difference between the group means, and the difference in variances can hardly reflect a correlation between variance and mean. It might arise from non-random sampling, since the subjects are volunteers, or it could be due to chance, since F is scarcely beyond the 5% level. As an exercise, test the difference between the means (i) without assuming $\sigma_1 = \sigma_2$, (ii) making this assumption. Ans. (i) $t' = 1.31$, $t_{0.05} = 2.17$, (ii) $t = 1.19$, $t_{0.05} = 2.048$. There is no difference in the conclusions.

EXAMPLE 4.15.3—In the preceding example, show that 95% confidence limits for $\mu_1 - \mu_2$ are -0.58 and 2.34 if we do not assume $\sigma_1 = \sigma_2$, and -0.63 and 2.39 if this assumption is made.

EXAMPLE 4.15.4—If you wanted to test the null hypothesis $\sigma_1 = \sigma_2$ from the data in table 4.14.1, would you use a one-tailed or a two-tailed test?

REFERENCES

1. W. J. YODEN and H. P. BEALE. *Contr. Boyce Thompson Inst.*, 6, 437 (1934).
2. L. C. GROVE. *Iowa Agric. Exp. Sta. Bul.*, 253 (1939).
3. H. H. MITCHELL, W. BURROUGHS, and J. R. BEADLES. *J. Nutrition*, 11:257 (1936).
4. E. W. CRAMPTON. *J. Nutrition*, 7:305 (1934).
5. W. R. BRENNEMAN. Personal communication.
6. O. W. PARK. *Iowa Agric. Exp. Sta. Bul.*, 151 (1932).
7. H. L. DEAN and R. H. WALKER. *J. Amer. Soc. Agron.*, 27:433 (1935).
8. S. N. SMITH. *J. Amer. Soc. Agron.*, 26:792 (1934).
9. P. B. PEARSON and H. R. CATCHPOLE. *Amer. J. Physiol.*, 115:90 (1936).
10. P. P. SWANSON and A. H. SMITH. *J. Biol. Chem.*, 97:745 (1932).
11. W. G. COCHRAN and G. M. COX. *Experimental Designs*. Wiley, New York, 2nd ed. (1957).
12. R. A. FISHER. *The Design of Experiments*, 7th ed. Oliver and Boyd, Edinburgh (1960).
13. D. R. COX. *Planning of Experiments*. Wiley, New York (1958).
14. G. W. CORNER. *The Hormones in Human Reproduction*. Princeton University Press (1943).
15. F. S. CHAPIN. *Experimental Designs in Sociological Research*. Harper, New York (1947).

16. W. V. BEHRENS. *Landwirtschaftliche Jahrbücher*, 68:807 (1929)
17. R. A. FISHER and F. YATES. *Statistical Tables*, 5th ed. Tables VI, VI₁ and VI₂ Oliver and Boyd, Edinburgh (1957)
18. A. A. ASPIN. *Biometrika*, 36:290 (1949).
19. W. G. COCHRAN. *Biometrics*, 20:191 (1964).
20. R. A. FISHER. *Proc. Int. Math. Conf. Toronto*, 805 (1924).
21. J. NEYMAN, K. IWASKIEWICZ, and ST KOŁODZIEJ(СЯК). *J. R. Statist. Soc.*, 2:114 (1935).
22. W. H. TRICKETT, B. L. WELCH, and G. S. JAMES. *Biometrika*, 43:203 (1956)

Shortcut and non-parametric methods

5.1—Introduction. In the preceding chapter you learned how to compare the means of two samples: paired or independent. The present chapter takes up several topics related to the same problem. For some years there has been continued activity in developing rapid and easy methods for dealing with samples from normal populations. In small samples, we saw that the range, as a substitute for the sample standard deviation, has remarkably high efficiency as compared to s . In section 5.2 a method will be described for comparing the means of two samples, using the range in place of s . Often this test, which is quickly made, leads to definite conclusions, so that there is no necessity to compute Student's t . This range test may also be employed as a rough check when there is doubt whether t has been computed correctly.

To this point the normal distribution has been taken as the source of most of our sampling. Fortunately, the statistical methods described are also effective for moderately abnormal populations. But there is much current interest in finding methods that work well for a wide variety of populations. Such methods, sometimes called *distribution-free* methods, are needed when sampling from populations that are far from normal. They are useful also, particularly in exploratory research, when the investigator does not know much about the type of distribution being sampled. The best-known procedures of this type are described in sections 5.3 to 5.7.

5.2—The t -test based on range. Lord (3) has developed an alternative to the t -test in which the range replaces $s_{\bar{x}}$ in the denominator of t . This test is used in the same way as t for testing a hypothesis or making interval estimates. Pillai (4) has shown that for interval estimates the efficiency of this procedure relative to t stays above 95% in samples up to $n = 20$. Like t , the range test assumes a normal distribution. It has become popular, particularly in industrial work.

Table A 7 (i) applies to single samples or to a set of differences obtained from paired samples. The entries are the values of $(\bar{X} - \mu)/w$, where w denotes the range of the sample. This ratio will be called t_w , since it plays the role of t .

For an illustration of the setting of confidence intervals by means of Lord's table, we use the vitamin C data from chapter 2. The sample values were 16, 22, 21, 20, 23, 21, 19, 15, 13, 23, 17, 20, 29, 18, 22, 16, 25, with $\bar{X} = 20$. We find $w = 29 - 13 = 16$ mg./100 gm., with $n = 17$. Table A 7 (i) has the entry 0.144 in the column headed 0.05 and the row for $n = 17$. The probability that $|t_w| \leq 0.144$ is 0.95 in random samples of $n = 17$ from a normally distributed population. The 95% confidence interval for μ is fixed by the inequalities

$$\bar{X} - t_w w \leq \mu \leq \bar{X} + t_w w$$

Substituting the vitamin C data,

$$20 - (0.144)(16) \leq \mu \leq 20 + (0.144)(16) \\ 17.7 \leq \mu \leq 22.3 \text{ mg./100 gm.}$$

This is to be compared with the slightly narrower interval $17.95 \leq \mu \leq 22.05$ based on s .

The test of a null hypothesis by means of t_w is illustrated by the paired samples in chapter 4 showing the numbers of lesions on the two halves of tobacco leaves under two preparations of virus. The eight differences between the halves were 13, 3, 4, 6, -1, 1, 5, 1. Here the mean difference $\bar{D} = 4$, while $w = 14$ and $n = 8$. For the null hypothesis that the two preparations produce on the average equal numbers of lesions,

$$t_w = \frac{\bar{D}}{w} = \frac{4}{14} = 0.286,$$

which is practically at the 5% level (0.288). The ordinary t -test gave a significance probability of about 4%.

Table A 7 (ii) applies to two independent samples of equal size. The mean of the two ranges, $\bar{w} = (\bar{w}_1 + \bar{w}_2)/2$, replaces the w of the preceding paragraphs and $\bar{X}_1 - \bar{X}_2$ takes the place of \bar{D} .

The test of significance will be applied to the numbers of worms found in two samples of five rats, one sample treated previously by a worm killer.

TABLE 5.2.1
NUMBER OF WORMS PER RAT

	Treated	Untreated
	123	378
	143	275
	192	412
	40	265
	259	286
Means, \bar{X}	151.4	323.2
Ranges, w	219	147

We have $\bar{X}_2 - \bar{X}_1 = 171.8$ and $\bar{w} = (219 + 147)/2 = 183$. From this, $t_w = (\bar{X}_2 - \bar{X}_1)/\bar{w} = 171.8/183 = 0.939$, which is beyond the 1% point, 0.896, shown in table A 7 (ii) for $n = 5$.

To find 95% confidence limits for the reduction in number of worms per rat due to the treatment, we use the formula

$$\begin{aligned} (\bar{X}_2 - \bar{X}_1) - t_w \bar{w} &\leq \mu_2 - \mu_1 \leq (\bar{X}_2 - \bar{X}_1) + t_w \bar{w} \\ 171.8 - (0.613)(183) &\leq \mu_2 - \mu_1 \leq 171.8 + (0.613)(183) \\ 60 &\leq \mu_2 - \mu_1 \leq 284 \end{aligned}$$

The confidence interval is wide, owing both to the small sample sizes and the high variability from rat to rat. Student's t , used in example 4.9.3 for these data, gave closely similar results both for the significance level and the confidence limits.

For two independent samples of unequal sizes, Moore (1) has given tables for the 10%, 5%, 2%, and 1% levels of Lord's test to cover all cases in which the sample sizes n_1 and n_2 are both 20 or less.

The range method can also be used when the sample size exceeds 20. With two samples each of size 24, for example, each sample may be divided at random into two groups of size 12. The range is found for each group, and the average of the four ranges is taken. Lord (3) gives the necessary tables. This device keeps the efficiency of the range test high for samples greater than 20, though the calculation takes a little longer.

To summarize, the range test is convenient for normal samples if a 5% to 10% loss in information can be tolerated. It is much used when many routine tests of significance or calculations of confidence limits have to be made. It is more sensitive than t to skewness in the population and to the appearance of gross errors.

EXAMPLE 5.2.1—In a previous example the differences in the serum albumen found by two methods A and B in eight blood samples were 0.6, 0.7, 0.8, 0.9, 0.3, 0.5, -0.5, 1.3 gm. per 100 ml. Apply the range method to test the null hypothesis that there is no consistent difference in the amount of serum albumen found by the two methods. Ans $t_w = 0.32$, $P < 0.05$.

EXAMPLE 5.2.2—In this example, given by Lord (3), the data are the times taken for an aqueous solution of glycerol to fall between two fixed marks. In five independent determinations in a viscometer, these times were 103.5, 104.1, 102.7, 103.2, and 102.6 seconds. For satisfactory calibration of the viscometer, the mean time should be accurate to within $\pm 1/2$ sec., apart from a 1-in-20 chance. By finding the half-width of the 95% confidence interval for μ by (i) the t_w method, and (ii) the t method, verify whether this requirement is satisfied. Ans No. Both methods give ± 0.76 for the half-width.

EXAMPLE 5.2.3—In 15 kernels of corn the crushing resistance of the kernels, in pounds, ranged from 25 to 65 with a mean of 43.0. Another sample of 15 kernels, harvested at a different stage, ranged from 29 to 67 with a mean of 48.0. Test whether the difference between the means is significant. Ans No, $t_w = 0.128$. Note that since the ranges of the two samples indicate much overlap, one could guess that the test will not show a significant difference.

5.3—Median, percentiles, and order statistics. The median of a population has the property that half the values in the population exceed it and half fall short of it. To estimate the median from a sample, arrange the observations in increasing order. When the sample values are arranged in this way, they are often called the 1st, 2nd, 3rd . . . *order statistics*. If the sample size n is odd, the sample median is the middle term in this array. For example, the median of the observations 5, 1, 8, 3, 4 is 4. In general, (n odd) the median is the order statistic whose number is $(n + 1)/2$. With n even, there is no middle term, and the median is defined as the average of the order statistics whose numbers are $n/2$ and $(n + 2)/2$. The median of the observations 1, 3, 4, 5, 7, 8 is 4.5.

Like the mean, the median is a measure of the middle of a distribution. If the distribution is symmetrical about its mean, the mean and the median coincide. With highly skewed distributions like that of income per family or annual sales of firms, the median is often reported, because it seems to represent people's concept of an average better than the mean. This point can be illustrated with small samples. As we saw, the median of the observations 1, 3, 4, 5, 8 is 4, while the mean is 4.2. If the sample values become 1, 3, 4, 5, 24, where the 24 simulates the introduction of a wealthy family or a large firm, the median is still 4, but the mean is 7.4. Four of the five sample values now fall short of the mean, while only one exceeds it. Similarly, in the distribution of incomes per family in a country, it is not unusual to find that 65% of families have incomes below the mean, with only 35% above it. In this sense, the mean does not seem a good indicator of the middle of the distribution. Further, the sample median in our small sample is still 4 even if we do not know the value of the highest observation, but merely that it is very large. With this sample, the mean cannot be calculated at all.

The calculation of the median from a large sample is illustrated from the data in table 5.3.1. This shows for 179 records of cows, the number of days between calving and the resumption of the oestrus cycle (16). Many of the records are repeated observations from successive calvings of the same cow. This raises doubts about the conclusions drawn, but the data are intended merely for illustration.

TABLE 5.3.1
DISTRIBUTION OF NUMBER OF DAYS FROM CALVING TO FIRST SUBSEQUENT OESTRUS
FOR A HOLSTEIN-FRIESIAN HERD IN WISCONSIN

Class limits (days)	0 5– 20 5	20 5– 40 5	40 5– 60 5	60 5– 80 5	80 5– 100 5	100 5– 120 5	120 5– 140 5	140 5– 160 5	160 5– 180 5	180 5– 200 5	200 5– 220 5
Frequency	8	33	50	32	15	20	11	6	2	1	1
Cumulative frequency	8	41	91	123	138	158	169	175	177	178	179

The frequency rises to a peak in the class from 40.5 days to 60.5 days. The day corresponding to the greatest frequency was called the *mode* by Karl Pearson. There is a secondary mode in the class from 100.5 to 120.5 days. This *bimodal* feature, as well as the skewness, emphasizes the non-normality of the distribution.

Since $n = 179$, the sample median is the order statistic that is 90th from the bottom. To find this, cumulate the frequencies as shown in the table until a cumulated frequency higher than 90 is reached—in this case 91. It is clear that the median is very close to the top of the 40.5–60.5 days class. The median is found by interpolation. Assuming that the 50 observations in this class are evenly distributed between 40.5 and 60.5 days, the median is 49/50 along the interval from 40.5 days to 60.5 days. The general formula is

$$M = X_L + \frac{gI}{f}, \quad (5.3.1)$$

where

X_L = value of X at lower limit of the class containing the median
= 40.5 days

g = order statistic number of the median *minus* cumulative frequency up to the upper limit of previous class = $90 - 41 = 49$

I = class interval = 20 days

f = frequency in class containing the median = 50

This gives

$$M = \text{Median} = 40.5 + \frac{(49)(20)}{50} = 60 \text{ days}$$

The mean of the distribution turns out to be 69.9 days, considerably higher than the median because of the long positive tail.

In large samples of size n from a normal distribution (6), the sample median becomes normally distributed about the population median with standard error $1.253\sigma/\sqrt{n}$. For this distribution, in which the sample mean and median are estimates of the same quantity, the median is less accurate than the mean. As we have stated, however, the chief application of the median lies in non-normal distributions.

There is a simple method of calculating confidence limits for the population median that is valid for any continuous distribution. Two of the order statistics serve as the upper and lower confidence limits. These are the order statistics whose numbers are, approximately (7),

$$\frac{(n+1)}{2} \pm \frac{z\sqrt{n}}{2}, \quad (5.3.2)$$

where z is the normal deviate corresponding to the desired confidence probability. For the sample of cows, using 95% confidence probability, $z \doteq 2$ and these numbers are $90 \pm \sqrt{179} = 77$ and 103. The 95% confidence limits are the numbers of days corresponding to the 77th and the 103rd order statistics. The actual numbers of days are found by adapting formula 5.3.1 for the median.

$$\text{For } 77: \text{No. of days} = 40.5 + \frac{(36)(20)}{50} = 55 \text{ days}$$

$$\text{For } 103: \text{No. of days} = 60.5 + \frac{(12)(20)}{32} = 68 \text{ days}$$

The population median is between 55 and 68 days unless this is one of those unusual samples that occur about once in twenty trials. The reasoning behind this method of finding confidence limits is essentially that by which confidence limits were found for the binomial in chapter 1. Formula 5.3.2 for finding the two-order statistics is a large-sample approximation, but is adequate for practical purposes down to $n = 25$.

In reporting on frequency distributions from large samples, investigators often quote *percentiles* of the distributions. The 90th percentile of a distribution of students' I.Q. scores is the I.Q. value such that 90% of the students fall short of it and only 10% exceed it.

In estimating percentiles, a useful result (7) is that in any continuous frequency distribution the P th percentile is estimated by the order statistic whose number is $(n + 1)P/100$. For the 179 cows, the 90th percentile is estimated by order statistic whose number is $i = (180)(90)/100 = 162$. By again using formula 5.3.1, the number of days corresponding to the 162nd order statistic is found as

$$120.5 + (4)(20)/11 = 128 \text{ days}$$

EXAMPLE 5.3.1—From a sample whose values are 8, 9, 2, 7, 3, 12, 15, estimate (i) the median, (ii) the *lower quartile* of the population (the lower quartile is the 25th percentile, having one-quarter of the population below it and three-quarters above), (iii) the 80th percentile. Ans (i) 8, (ii) 3, (iii) 13.2. For the 80th percentile, the number of the order statistic is 6.4. Since the 6th and 7th order statistics have values 12 and 15, respectively, linear interpolation gives 13.2 for the 6.4th order statistic. Note that from this small sample we cannot estimate the 90th percentile, beyond saying that our estimate exceeds 15.

5.4—The sign test. Often there is no scale for measuring a character, yet one believes that he can distinguish grades of merit. The animal husbandman, for example, judges body conformation, ranking the individuals from high to low, then assigning *ranks* 1, 2, . . . n . In the same way, the foods expert arrays preparations according to flavor or palatability. If rankings of a set of individuals or treatments are made by a random sample of judges, inferences can be made about the ranking in the population from which the sample of judges was drawn; this despite the fact that the parameters of the distributions cannot be written down.

First consider the rankings of two products by each of m judges. As an example, $m = 8$ judges ranked patties of ground beef which had been stored for 8 months at two temperatures in home freezers (17). Flavor was the basis of the ranking. Eight of the patties, one for each judge, were kept at $0^{\circ}\text{F}.$; the second sample of 8 were in a freezer whose temperature fluctuated between 0° and $15^{\circ}\text{F}.$ The rankings are shown in table 5.4.1.

TABLE 5.4.1
RANKINGS OF THE FLAVOR OF PAIRS OF PATTIES OF GROUND BEEF
(Eight judges. Rank 1 is high; rank 2, low)

Judge	Sample 1 $0^{\circ}\text{F}.$	Sample 2 Fluctuated
A	1	2
B	1	2
C	2	1
D	1	2
E	1	2
F	1	2
G	1	2
H	1	2

There are two null hypotheses that might be considered for these data. One is that the fluctuation in temperature produces no detectable difference in flavor. (If this hypothesis is true, however, one would expect some of the judges to report that their two patties taste alike and to be unwilling to rank them.) A second null hypothesis is that there is a difference in flavor, and that in the population from which the judges were drawn, half the members prefer the patties kept at $0^{\circ}\text{F}.$ and half prefer the other patties. Both hypotheses have the same consequence as regards the experimental data—namely, that for any judge in the sample, the probability is $1/2$ that the $0^{\circ}\text{F}.$ patty will be ranked 1. The reasons for this statement are different in the two cases. Under the first null hypothesis, the probability is $1/2$ because the rankings are arbitrary; under the second, because any judge drawn into the sample has a probability $1/2$ of being a judge who prefers the $0^{\circ}\text{F}.$ patty.

In the sample, 7 out of 8 judges preferred the $0^{\circ}\text{F}.$ patty. On either null hypothesis, we expect 4 out of 8. The problem of testing this hypothesis is exactly the same as that for which the χ^2 test was introduced in sections 1.10, 1.11 of chapter 1. From the general formula in section 1.12,

$$\chi^2 = \frac{(7 - 4)^2}{4} + \frac{(1 - 4)^2}{4} = 4.5$$

When testing the null hypothesis that the probability is $1/2$, a slightly simpler version of this formula is

$$\chi^2 = \frac{(a - b)^2}{n} = \frac{(7 - 1)^2}{8} = 4.5$$

where a and b are the observed numbers in the two classes (0°F. and Fluctuated).

Since the sample is small, we introduce a *correction for continuity*, described in section 8.6, and compute χ^2 as

$$\chi^2 = \frac{(|a - b| - 1)^2}{n} = \frac{(6 - 1)^2}{8} = 3.12, P = 0.078$$

The expression $|a - b| - 1$ means that we reduce the absolute value of $(a - b)$ by 1 before squaring. The test indicates non-significance, though the decision is close.

In this example we used the χ^2 test, in place of the t -test for paired samples, because the individual observations, instead of being distributed normally, take only the values 1 or 2, so that the differences within a pair are either +1 or -1. The same test is often used with continuous or discrete data, either because the investigator wishes to avoid the assumption of normality or as a quick substitute for the t -test. The procedure is known as the *sign test* (8), because the differences between the members of a pair are replaced by their signs (+ or -), the size of the difference being ignored. In the formula for χ^2 , a and b are the numbers of + and - signs, respectively. Any zero difference is omitted from the test, so that $n = a + b$.

When the sign test is applied to a variate X that has a continuous or discrete distribution, the null hypothesis is that X has the same distribution under the two treatments. But the null hypothesis does not need to specify the shape of this distribution. In the t -test, on the other hand, the null hypothesis assumes normality and specifies that the parameter μ (the mean) is equal for the two treatments. For this reason the t -test is sometimes called a *parametric* test, while the sign test is called *non-parametric*. Similarly, the median and other order statistics are non-parametric estimates, since they estimate percentiles of any continuous distribution without our requiring to define the shape of the distribution specifically by means of parameters.

In sampling from normal distributions the efficiency of the sign test relative to the t -test is about 65%. This statement implies that if the null hypothesis is false, so that the means of the two populations differ by an amount δ , a sign test based on 18 pairs and a t -test based on 12 pairs have about the same probability of detecting this by finding a significant difference. The sign test saves time at the expense of a loss of sensitivity in the test.

For numbers of pairs up to 20, table A 8 (p 554), intended for quick reference, shows the *smaller* number of like signs required for significance

at the 1%, 5%, and 10% levels. For instance, with 18 pairs, we must have 4 or less of one sign and 14 or more of the other sign in order to attain 5% significance. This table was computed not from the χ^2 approximation but from the exact binomial distribution. Since this distribution is discontinuous, we cannot find sample results that lie precisely at the 5% level. The significance probabilities, which are often substantially lower than the nominal significance levels, are shown in parentheses in table A 8. The finding of 4 negative and 14 positive signs out of 18 represents a significance probability of 0.031 instead of the nominal 0.05. For one-tailed tests these probabilities should be halved.

EXAMPLE 5.4.1—On being presented with a choice between two sweets, differing in color but otherwise identical, 15 out of 20 children chose color B. Test whether this is evidence of a general preference for B (i) by χ^2 , (ii) by reference to table A 8. Do the results agree?

EXAMPLE 5.4.2—Two ice creams were made with different flavors but otherwise similar. A panel of 6 expert dairy industry men all ranked flavor A as preferred. Is this statistical evidence that the consuming public will prefer A?

EXAMPLE 5.4.3—To illustrate the difference between the sign test and the t -test in extreme situations, consider the two samples, each of 9 pairs, in which the actual differences are as follows. Sample I: -1, 1, 2, 3, 4, 4, 6, 7, 10. Sample II: 1, 1, 2, 3, 4, 4, 6, 7, -10. In both samples the sign test indicates significance at the 5% level, with $P = 0.039$ from table A 8. In sample I, in which the negative sign occurs for the smallest difference, we find $t = 3.618$, with 8 *d.f.*, the significance probability being 0.007. In sample II, where the largest difference is the one with the negative sign, $t = 1.125$, with $P = 0.294$. Verify that Lord's test shows $t_w = 0.364$ for sample I and 0.118 for sample II, and gives verdicts in good agreement with the t -test. When the aberrant signs represent extreme observations the sign test and the t -test do not agree well. This does not necessarily mean that the sign test is at fault; if the extreme observation were caused by an undetected gross error, the verdict of the t -test might be misleading.

5.5—Non-parametric methods: ranking of differences between measurements. The *signed rank* test, due to Wilcoxon (2), is another substitute for the t -test in paired samples. First, the absolute values of the differences (ignoring signs) are ranked, the smallest difference being assigned rank 1. Then the signs are restored to the rankings. The method is illustrated from an experiment by Collins *et al.* (9). One member of a pair of corn seedlings was treated by a small electric current, the other being untreated. After a period of growth, the differences in elongation (treated-untreated) are shown for each of ten pairs.

In table 5.5.1 the ranks with negative signs total 15 and those with positive signs total 40. The test criterion is the *smaller* of these totals, in this case, 15. The ranks with the less frequent sign will usually, though not always, give the smaller rank total. This number, sign ignored, is referred to table A 9. For 10 pairs a rank sum ≤ 8 is required for rejection at the 5% level. Since $15 > 8$, the data support the null hypothesis that elongation was unaffected by the electric current treatment.

The null hypothesis in this test is that the frequency distribution of the original measurements is the same for the treated and untreated mem-

TABLE 5.5.1
EXAMPLE OF WILCOXON'S SIGNED RANK TEST
(Differences in elongation of treated and untreated seedlings)

Pair	Difference (mm.)	Signed Rank
1	6.0	5
2	1.3	1
3	10.2	7
4	23.9	10
5	3.1	3
6	6.8	6
7	- 1.5	-2
8	-14.7	-9
9	- 3.3	-4
10	11.1	8

bers of a pair, but as in the sign test the shape of this frequency distribution need not be specified. A consequence of this null hypothesis is that each rank is equally likely to have a + or a - sign. The frequency distribution of the smaller rank sum was worked out by the rules of probability as described by Wilcoxon (2). Since this distribution is discontinuous, the significance probabilities for the entries in table A 9 are not exactly 5% and 1%, but are close enough for practical purposes.

If the two or more differences are equal, it is often sufficiently accurate to assign to each of the ties the average of the ranks that would be assigned to the group. Thus, if two differences are tied in the fifth and sixth positions, assign rank 5 1/2 to each of them.

If the number of pairs n exceeds 16, calculate the approximate normal deviate

$$Z = (|\mu - T| - \frac{1}{2})/\sigma$$

where T is the smaller rank sum, and

$$\mu = n(n+1)/4 \quad : \quad \sigma = \sqrt{(2n+1)\mu/6}$$

The number $-1/2$ is a correction for continuity. As usual, $Z > 1.96$ signifies rejection at the 5% level.

EXAMPLE 5.5.1 - From two J-shaped populations distributed like chi-square with $d.f. = 1$ (figure 1.13.1), two samples of $n = 10$ were drawn and paired at random:

Sample 1	1.98	3.30	5.91	1.05	1.01	1.44	3.42	2.17	1.37	1.13
Sample 2	0.33	0.11	0.04	0.24	1.56	0.42	0.00	0.22	0.82	2.54
Difference	1.65	3.19	5.87	0.81	-0.55	1.02	3.42	1.95	0.55	-1.41
Rank	6	8	10	3	-1.5	4	9	7	1.5	-5

The difference between the population means was 1. Apply the signed rank test. Ans. The smallest two absolute differences are tied, so each is assigned the rank $(1+2)/2 = 1.5$.

The sum of the negative ranks is 6.5, between the critical sums, 3 and 8, in table A 9. H_0 is rejected with $P = 0.04$, approximately.

EXAMPLE 5.5.2—If you had not known that the differences in the foregoing example were from a non-normal population, you would doubtless have applied the t -test. Would you have drawn any different conclusions? Ans. $t = 2.48$, $P = 0.04$.

EXAMPLE 5.5.3—Apply the signed rank test to samples I and II of example 5.4.3. Verify that the results agree with those given by the t -test and not with those given by the sign test. Is this what you would expect?

EXAMPLE 5.5.4—For 16 pairs, table A 9 states that the 5% level of the smaller rank sum is 29, the exact probability being 0.053. Check the normal approximation in this case by showing that $\mu = 68$, $\sigma = 19.34$, so that for $T = 29$ the value of Z is 1.99, corresponding to a significance probability of 0.047.

5.6—Non-parametric methods: ranking for unpaired measurements.

Turning now to the two-sample problems of chapter 4, we consider ranking as a non-parametric method for random samples of measurements which do not conform to the usual models. This test was also developed by Wilcoxon (2), though it is sometimes called the Mann-Whitney test (11). A table due to White (12) applies to unequal group sizes as well as equal. All observations in both groups are put into a single array, care being taken to tag the numbers of each group so that they can be distinguished. Ranks are then assigned to the combined array. Finally, the smaller sum of ranks, T , is referred to table A 10 to determine significance. Note that *small* values of T cause rejection.

An example is drawn from the Corn Borer project in Boone County, Iowa. It is well established that, in an attacked field, more eggs are deposited on tall plants than on short ones. For illustration we took records of numbers of eggs found in 20 plants in a rather uniform field. The plants were in 2 randomly selected sites, 10 plants each. Table 5.6.1 contains the egg counts.

TABLE 5.6.1
NUMBER OF CORN BORER EGGS ON CORN PLANTS, BOONE COUNTY, IOWA, 1950

Height of Plant	Number of Eggs									
Less than 23"	0	14	18	0	31	0	0	0	11	0
More than 23"	37	42	12	32	105	84	15	47	51	65

In years such as 1950 the frequency distribution of number of eggs tends to be J-shaped rather than normal. At the low end, many plants have no eggs, but there is also a group of heavily infested plants. Normal theory cannot be relied upon to yield correct inferences from small samples.

For convenience in assigning ranks, the counts were rearranged in increasing order (table 5.6.2). The counts for the tall plants are in bold-

TABLE 5.6.2
EGG COUNTS ARRANGED IN INCREASING ORDER, WITH RANKS
(**Boldface type indicates counts on plants 23" or more**)

Count	0,	0,	0,	0,	0,	0,	11,	12,	14,	15,	18,	31
Rank	3½,	3½,	3½,	3½,	3½,	3½,	7,	8,	9,	10,	11,	12

face type. The eight highest counts are omitted, since they are all on tall plants and it is clear that the small plants give the smaller rank sum.

By the rule suggested for tied ranks, the six ties are given the rank $3\frac{1}{2}$, this being the average of the numbers 1 to 6. In this instance the average is not necessary, since all the tied ranks belong to one group; the sum of the six ranks, 21, is all that we need. But if the tied counts were in both groups, averaging would be required.

The next step is to add the n_1 rank numbers in the group (plants less than 23 in.) that has the smaller sum.

$$T = 21 + 7 + 9 + 11 + 12 = 60$$

This sum is referred to table A 10 with $n_1 = n_2 = 10$. Since T is less than $T_{0.01} = 71$, the null hypothesis is rejected with $P \leq 0.01$. The anticipated conclusion is that plant height affects the number of eggs deposited.

When the samples are of unequal sizes n_1, n_2 , an extra step is required. First, find the total T_1 of the ranks for the sample that has the smaller size, say n_1 . Compute $T_2 = n_1(n_1 + n_2 + 1) - T_1$. Then T , which is referred to table A 10, is the smaller of T_1 and T_2 . To illustrate, White quotes Wright's data (10) on the survival times, under anoxic conditions, of the peroneal nerves of 4 cats and 14 rabbits. For the cats, the times were 25, 33, 43, and 45 minutes; for the rabbits, 15, 16, 16, 17, 20, 22, 22, 23, 28, 28, 30, 30, 35, and 35 minutes. The ranks for the cats are 9, 14, 17, and 18, giving $T_1 = 58$. Hence, $T_2 = 4(19) - 58 = 18$, and is smaller than T_1 , so that $T = 18$. For $n_1 = 4, n_2 = 14$, the 5% level of T is 19. The mean survival time of the nerves is significantly higher for the cats than for the rabbits.

For values of n_1 and n_2 outside the limits of the table, calculate

$$Z = (|\mu - T| - \frac{1}{2})/\sigma,$$

where

$$\mu = n_1(n_1 + n_2 + 1)/2 \quad ; \quad \sigma = \sqrt{n_2\mu/6}$$

The approximate normal deviate Z is referred to the tables of the normal distribution to give the significance probability P .

Table A 10 was calculated from the assumption that if the null hypothesis is true, the n_1 ranks in the smaller sample are a random selection from the $(n_1 + n_2)$ ranks in the combined samples.

5.7—Comparison of rank and normal tests. When the t -test is used on non-normal data, two things happen. The significance probabilities are changed; the probability that t exceeds $t_{0.05}$ when the null hypothesis is true is no longer 0.50, but may be, say, 0.041 or 0.097. Secondly, the sensitivity or power of the test in finding a significant result when the null hypothesis is false is altered. Much of the work on non-parametric methods is motivated by a desire to find tests whose significance probabilities do not change and whose sensitivity relative to competing tests remains high when the data are non-normal.

With the rank tests, the significance levels remain the same for any continuous distribution, except that they are affected to some extent by ties, and by zeros in the signed rank test. In large normal samples, the rank tests have an efficiency of about 95% relative to the t -test (13), and in small normal samples, the signed rank test has been shown (14) to have an efficiency slightly higher than this. With non-normal data from a continuous distribution, the efficiency of the rank tests relative to t never falls below 86% in large samples and may be much greater than 100% for distributions that have long tails (13). Since they are relatively quickly made, the rank tests are highly useful for the investigator who is doubtful whether his data can be regarded as normal.

The beginner may wish to compute both the rank tests and the t -test for some of his data to see how they compare. Needless to say, the practice of quoting the test that agrees with one's predilections vitiates the whole technique.

As has been stated previously, most investigations, after the preliminary stages, are designed to estimate the sizes of differences rather than simply to test null hypotheses. The rank methods can furnish estimates and confidence limits for the difference between two treatments (see examples 5.8.1 and 5.8.2). The calculations require no assumption of normality, but are a little tedious. Some work has also been done in extending rank methods to the more complex types of data that we shall meet in later chapters, though the available techniques still fall short of the flexibility of the standard methods based on normality.

5.8—Scales with limited values. In some lines of work the scales of measurement are restricted to a small number of values, perhaps to 0, 1, 2 or 1, 2, 3, 4, 5. Investigators are sometimes puzzled as to how to test the differences between two treatments in this case, because the data do not look normal, while rank methods usually involve a substantial number of zeros and ties. We suggest that the ordinary t -test be used, with the inclusion of a correction for continuity. To illustrate, consider a paired test in which the original data are on a 0, 1, 2 scale. The differences between the members of a pair can then assume only the values 2, 1, 0, -1, and -2.

With 12 pairs, suppose that the differences D between two treatments A and B are 2, 2, 2, 1, 1, 1, 0, 0, 0, 0, -1, -1. Then $\Sigma D = 7$. There is a

test, called Fisher's *randomization test* (15), that requires no assumption about the form of the basic distribution of these differences. The argument used is that if there is no difference between A and B, each of the 12 differences is equally likely to be + or -. Thus, under the null hypothesis there are $2^{12} = 4,096$ possible sets of sample results. Since, however, +0 and -0 are the same, only $2^8 = 256$ need be examined. We then count how many samples have ΣD as great as or greater than 7, the observed ΣD . It is not hard to verify that 38 samples are of this kind if both positive and negative totals are counted so as to provide a two-tailed test. The significance probability is $38/256 = 0.148$. The null hypothesis is not rejected by the randomization test.

With this test the investigator must work out his own significance probability. From his writings it seems clear that Fisher did not intend the test for routine use, but merely to illustrate that a test can be made if A and B were assigned to the members of each pair by randomization.

For scales with limited numbers of values, numerous comparisons of the results of this test and the *t*-test show that they usually agree well enough for practical purposes. In the randomization test, however, the possible values of ΣD jump by 2's. Our observed ΣD is 7. We would have $\Sigma D = 9$ if only one 1 had a - sign, and $\Sigma D = 5$ if three 1's had a - sign. To apply the correction for continuity, we compute t_c as

$$t_c = \frac{|\Sigma D| - 1}{ns_B} = \frac{6}{(12)(0.313)} = 1.597,$$

where $s_B = 0.313$ is computed in the usual way. With 11 *df.*, *P* is 0.138, in good agreement with the randomization test. The denominator of t_c is the standard error of ΣD . This may be computed either as ns_B or as $\sqrt{ns_D}$.

In applying the correction for continuity, the rule is to find the next highest value of ΣD that the randomization set provides. The numerator of t_c is halfway between this value and the observed ΣD . The values of ΣD do not always jump by 2's.

With two independent samples of size *n* the randomization test assumes that on the null hypothesis the $(2n)$ observations have been divided at random into two samples of *n*. There are $(2n)!/(n!)^2$ cases. To apply the correction, find the next highest value of $\Sigma D_1 - \Sigma D_2$. If one sample has the values 2, 3, 3, 3 and the other has 0, 0, 0, 2, we have $\Sigma D_1 = 11$, $\Sigma D_2 = 2$, giving $\Sigma D_1 - \Sigma D_2 = 9$. The next highest value is 7, given by the case 2, 2, 3, 3 and 0, 0, 0, 3. Hence, the numerator of t_c is 8. The general formula for t_c is

$$t_c = \frac{|\Sigma D_1 - \Sigma D_2| - c}{\sqrt{(ns_1^2 + ns_2^2)}}$$

with $2(n-1) d.f.$, where s_1^2 and s_2^2 are the sample variances and c is the size of the correction.

With small samples that show little overlap, as in this example, the randomization test is easily calculated and is recommended, because in such cases t_c tends to give too many significant results. With sample values of 2, 3, 3, 3 and 0, 0, 0, 2, the observed result is the most extreme of the $8!/(4!)^2$ cases. The randomization provides 4 cases like the observed one in a two-tailed test. P is therefore $4/70 = 0.057$. The reader may verify that $t_c = 3.58$, with 6 $d.f.$ and P near 0.01.

EXAMPLE 5 8 1—In Wright's data, p. 131, show that if the survival time for each cat is reduced by 2 minutes, the value of T in the signed rank test becomes $18\frac{1}{2}$, while if the cat times are reduced by 3 minutes, $T = 21$. Show further that if 23 minutes are subtracted from each cat, we find $T = 20\frac{1}{2}$, while for 24 minutes, $T = 19$. Since $T_{0.05} = 19$, any hypothesis which states that the average survival time of cats exceeds that of rabbits by a figure between 3 and 23 minutes is accepted in a 5% test. The limits 3 and 23 minutes are 95% confidence limits as found from the rank sum test.

EXAMPLE 5 8 2—In a two-sample comparison, the estimate of the difference between the two populations appropriate to the use of ranks is the median of the differences $X_i - Y_j$, where X_i and Y_j denote members of the first and second samples. In Wright's data, with $n_1 = 4$, $n_2 = 14$, there are 56 differences. Show that the median is 12.5. (You should be able to shortcut the work.)

EXAMPLE 5 8 3—In a paired two-sample test the ten values of the differences D were 3, 3, 2, 1, 1, 1, 1, 0, 0, -1. Show that the randomization test gives $P = 3/64 = 0.047$ while the value of t , corrected for continuity, is 2.457, corresponding to a P value of about 0.036.

REFERENCES

- 1 P. G. MOORE *Biometrika*, 44 482 (1957)
- 2 F. WILCOXON *Biometrics Bul.*, 1 80 (1945)
- 3 E. LORD *Biometrika*, 34 56 (1947)
- 4 K. C. S. PILLAI *Ann. Math. Statist.*, 22 469 (1951)
- 5 C. M. HARRISON *Plant Physiol.*, 9 94 (1934)
- 6 M. G. KENDALL and A. STUART *The Advanced Theory of Statistics*. Vol. I, 2nd ed. Charles Griffin, London (1958).
- 7 A. M. MOOD and F. A. GRAYBILL *Introduction to the Theory of Statistics*, 2nd ed., p. 408. McGraw-Hill, New York (1963)
- 8 W. J. DIXON and A. M. MOOD *J. Amer. Statist. Ass.*, 41 557 (1946)
- 9 G. N. COLLINS, *et al.* *J. Agric. Res.*, 38 585 (1929)
- 10 E. B. WRIGHT *Amer. J. Physiol.*, 147 78 (1946)
- 11 H. B. MANN and D. R. WHITNEY *Ann. Math. Statist.*, 18 50 (1947)
- 12 C. WHITE *Biometrics*, 8 33 (1952)
- 13 J. L. HODGES and E. L. LEHMANN *Ann. Math. Statist.*, 27 324 (1956)
- 14 J. KLOTZ *Ann. Math. Statist.*, 34 624 (1963)
- 15 R. A. FISHER *The Design of Experiments*, 7th ed., p. 44. Oliver and Boyd, Edinburgh (1960)
- 16 A. B. CHAPMAN and L. E. CASIDA *J. Agric. Res.*, 54 417 (1937)
- 17 F. EHRENKRANTZ and H. ROBERTS *J. Home Econ.*, 44 441 (1952)

Regression

6.1—Introduction. In preceding chapters the problems considered have involved only a single measurement on each individual. In this chapter, attention is centered on the dependence of one variable Y on another variable X . In mathematics Y is called a *function* of X , but in statistics the term *regression* is generally used to describe the relationship. The growth curve of height is spoken of as the regression of height on age; in toxicology the lethal effects of a drug are described by the regression of per cent kill on the amount of the drug. The origin of the term regression will be explained in section 6.16. To distinguish the two variables in regression studies, Y is sometimes called the *dependent* and X the *independent* variable. These names are fairly appropriate in the toxicology example, in which we can think of the per cent kill Y as being caused by the amount of drug X , the amount itself being variable at the will of the investigator. They are less suitable though still used, for example, when Y is the weight of a man and X is his maximum girth.

Regression has many uses. Perhaps the objective is only to learn if Y does depend on X . Or, prediction of Y from X may be the goal. Some wish to determine the shape of the regression curve. Others are concerned with the error in Y in an experiment after adjustments have been made for the effect of a related variable X . An investigator has a theory about cause and effect, and employs regression to test this theory. To satisfy these various needs an extensive account of regression methods is necessary.

In the next two sections the calculations required in fitting a regression are introduced by a numerical example. The theoretical basis of these calculations and the useful applications of regression are taken up in subsequent sections.

6.2.—The regression of blood pressure on age. A project "The Nutritional Status of Population Groups" was set up by the Agricultural Experiment Stations of nine midwestern states. From the facts learned we have extracted data on systolic blood pressure among 58 women over 30 years of age, a random sample from a region near Ames, Iowa (1). For

present purposes, the ages are grouped into 10-year classes and the mean blood pressure calculated for each class. The results are in the first two columns of table 6.2.1.

TABLE 6.2.1
MEAN SYSTOLIC BLOOD PRESSURE OF 58 WOMEN IN 10-YEAR AGE CLASSES

Midpoint of Age Class X	Mean Blood Pressure Y	Deviations From Means x y		Squares x^2 y^2		Products xy
35	114	-20	-27	400	729	540
45	124	-10	-17	100	289	170
55	143	0	2	0	4	0
65	158	10	17	100	289	170
75	166	20	25	400	625	500
Sum	275	0	0	1,000	1,936	1,380
Mean	55					

Sample regression coefficient $b = \frac{\Sigma xy}{\Sigma x^2} = \frac{1,380}{1,000} = 1.38$ units of blood pressure per year

As in most regression problems, the first thing to do is to draw a graph, figure 6.2.1. The independent variable X is plotted along the horizontal axis. Each measure of the dependent Y is indicated by a black circle above the corresponding X . Clearly, the trend of blood pressure with age is upward and roughly linear.

The straight line drawn in the figure is the *sample regression of Y on X* . Its position is fixed by two results:

(i) It passes through the point $O'(\bar{X}, \bar{Y})$, the point determined by the mean of each sample. For the blood pressures this is the point (55, 141).

(ii) Its slope is at the rate of b units of Y per unit of X , where b is the *sample regression coefficient*. Writing $x = X - \bar{X}$ and $y = Y - \bar{Y}$, $b = \Sigma xy / \Sigma x^2$. The numerator of b is a new quantity—the sum of products of the deviations, x and y . In table 6.2.1 the individual values of x^2 have been obtained in the fifth column and those of xy in the seventh column. In section 6.3 a quicker method of calculating b will be given. For the blood pressures, $b = +1.38$, meaning that blood pressure increases on the average by 1.38 units per year of age.

The sample regression equation of Y on X is now written as

$$\begin{aligned} \hat{Y} &= \bar{Y} + bx, \\ \text{or, } \hat{y} &= bx, \end{aligned} \quad \left. \vphantom{\begin{aligned} \hat{Y} &= \bar{Y} + bx, \\ \hat{y} &= bx, \end{aligned}} \right\}$$

where \hat{Y} is the estimated value and \hat{y} the estimated deviation of Y corresponding to any x -deviation. If $x = 20$ years, $\hat{y} = (1.38)(20) = 27.6$ units of blood pressure

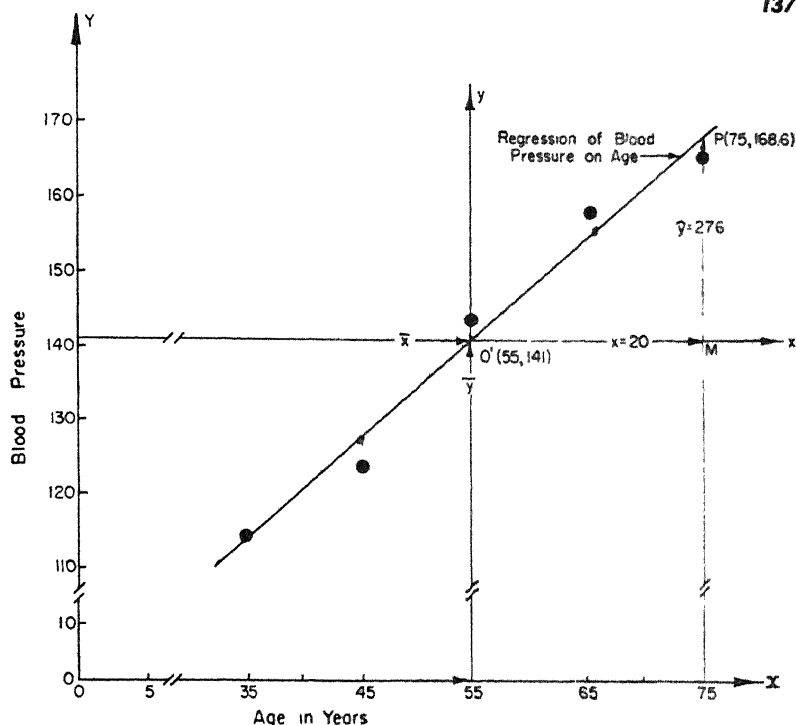


FIG 6.2.1— Sample regression of blood pressure on age. The broken lines indicate omission of the lower parts of the scales in order to clarify the relations in the parts occupied by the data.

This equation enables us to complete figure 6.2.1 by drawing the sample regression line. Lay off $O'M = 20$ years to the right of O' , then erect a perpendicular, $MP = 27.6$ units of blood pressure. The line $O'P$ then has the slope, 1.38 units of blood pressure per year.

In terms of the original units, the sample regression equation is

$$\hat{Y} - \bar{Y} = b(X - \bar{X})$$

For the blood pressures, this becomes

$$\hat{Y} - 141 = 1.38(X - 55)$$

or

$$\begin{aligned}\hat{Y} &= 141 + 1.38(X - 55) \\ &= 65.1 + 1.38X\end{aligned}$$

If $X = 75$ is entered in this equation, \hat{Y} becomes $65.1 + (1.38)(75) = 168.6$ units of blood pressure. The corresponding point, $(75, 168.6)$, is shown as P in the figure.

We can now compare the sample points with the corresponding \hat{Y} to

get measures of the *goodness of fit* of the line to the data. Each X is substituted in the regression equation and \hat{Y} calculated. The five results are recorded in table 6.2.2. The *deviations from regression*, $Y - \hat{Y} = d_{y \cdot x}$, measure the failure of the line to fit the data. In this sample, 45-year-old women had below average blood pressure and 65-year-olds had an excess.

TABLE 6.2.2
CALCULATION OF \hat{Y} AND DEVIATIONS FROM REGRESSION, $d_{y \cdot x} = Y - \hat{Y}$
(Blood pressure data)

Midpoint of Age Class X	Mean Blood Pressure Y	Estimated Blood Pressure \hat{Y}	Deviation From Regression $Y - \hat{Y} = d_{y \cdot x}$	Square of Deviation $d_{y \cdot x}^2$
35	114	113.4	0.6	0.36
45	124	127.2	-3.2	10.24
55	143	141.0	2.0	4.00
65	158	154.8	3.2	10.24
75	166	168.6	-2.6	6.76
Sum			$\Sigma d_{y \cdot x} = 0.0$	$\Sigma d_{y \cdot x}^2 = 31.60$

The sum of squares of deviations, $\Sigma d_{y \cdot x}^2 = 31.60$, is the basis for an estimate of error in fitting the line. The corresponding degrees of freedom are $n - 2 = 3$. We have then,

$$s_{y \cdot x}^2 = \Sigma d_{y \cdot x}^2 / (n - 2) = 10.53,$$

where $s_{y \cdot x}^2$ is the *mean square deviation from regression*. The resulting *sample standard deviation from regression*,

$$s_{y \cdot x} = \sqrt{s_{y \cdot x}^2} = 3.24 \text{ units of blood pressure,}$$

corresponds to s in single-variable problems. In particular, it furnishes a *sample standard deviation of the regression coefficient*,

$$s_b = s_{y \cdot x} / \sqrt{\Sigma x^2}$$

This is $3.24 / \sqrt{1,000} = 0.102$ units of blood pressure, with $(n - 2) = 3$ d.f.

A test of significance of b is given by

$$t = b/s_b, \text{ d.f.} = n - 2$$

Applying this to the blood pressures,

$$t = 1.38/0.102 = 13.5^{**}, \text{ d.f.} = 3$$

Note: It is often convenient to denote significance by asterisks. A single one indicates probabilities between 0.05 and 0.01; two indicate probabilities equal to or less than 0.01.

Often there is little interest in the individual $d_{y \cdot x}$ of table 6.2.2. If so, $\Sigma d_{y \cdot x}^2$ may be calculated directly by the formula,

$$\Sigma d_{y \cdot x}^2 = \Sigma y^2 - [(\Sigma xy)^2 / \Sigma x^2]$$

Substituting the blood pressure data from table 6.2.1,

$$\Sigma d_{y \cdot x}^2 = 1,936 - [(1,380)^2 / 1,000] = 31.60$$

as before.

EXAMPLE 6.2.1—Following are measurements on heights of soybean plants in a field, a different random selection each week (2)

Age in weeks	1	2	3	4	5	6	7
Height in centimeters	5	13	16	23	33	38	40

Verify these results. $\bar{X} = 4$ weeks, $\bar{Y} = 24$ cms., $\Sigma x^2 = 28$, $\Sigma y^2 = 1,080$, $\Sigma xy = 172$ Compute the sample regression, $\hat{Y} = 6.143 X - 0.572$ centimeters

EXAMPLE 6.2.2—Plot on a graph the sample points for the soybean data, then construct the sample regression line. Do the points lie about equally above and below the line?

EXAMPLE 6.2.3—Calculate $s_b = 0.409$ cms./wk Set the 95% confidence interval for the population regression. Ans. 5.09 – 7.20 cms./wk Note that s_b has 5 df

EXAMPLE 6.2.4—The soybean data constitute a growth curve. Do you suppose the population growth curve is really straight? How would you design an experiment to get a growth curve of the blood pressure in Iowa women?

EXAMPLE 6.2.5—Eighteen samples of soil were prepared with varying amounts of inorganic phosphorus, X . Corn plants, grown in each soil, were harvested at the end of 38 days and analyzed for phosphorus content From this was estimated the plant-available phosphorus in the soil. Nine of the observations, adapted for ease of computation, are shown in this table.

Inorganic phosphorus in soil (ppm), X	1	4	5	9	13	11	23	23	28
Estimated plant-available phosphorus (ppm), Y	64	71	54	81	93	76	77	95	109

Calculate $b = 1.417$, $s_b = 0.395$, $t = 3.59^{**}$

6.3—Shortcut methods of computation in regression. Since regression computations are tedious, a calculating machine is almost essential. In fitting a regression, the following six basic quantities must be obtained:

$$n, \bar{X}, \bar{Y}, \Sigma x^2, \Sigma y^2, \Sigma xy$$

You already know shortcut methods of computing Σx^2 and Σy^2 without finding the individual deviations x and y . A similar method exists for finding Σxy , based on the algebraic identity

$$\Sigma xy = \Sigma (X - \bar{X})(Y - \bar{Y}) = \Sigma XY - (\Sigma X)(\Sigma Y)/n$$

Note that the correction term may be larger than ΣXY , making Σxy negative. This indicates a *downward sloping* regression line

In table 6.3.1 the regression of blood pressure on age has been recomputed using these shortcuts.

TABLE 6 3 1
MACHINE COMPUTATION OF A LINEAR REGRESSION

Age (years), X	35	45	55	65	75
Blood pressure (units), Y	114	124	143	158	166
$\Sigma X = 275$	$\Sigma Y = 705$	$n = 5$			
$\bar{X} = 55$	$\bar{Y} = 141$				
$\Sigma X^2 = 16,125$	$\Sigma Y^2 = 101,341$	$\Sigma XY = 40,155$			
$(\Sigma X)^2/n = 15,125$	$(\Sigma Y)^2/n = 99,405$	$(\Sigma X)(\Sigma Y)/n = 38,775$			
$\Sigma x^2 = 1,000$	$\Sigma y^2 = 1,936$	$\Sigma xy = 1,380$			
$b = \Sigma xy / \Sigma x^2 = 1,380/1,000 = 1.38$ units per year of age $\hat{Y} = \bar{Y} + b(\bar{x} - \bar{X})$ $= 141 + 1.38(X - 55) = 65.1 + 1.38X$					
$\Sigma d_y^2 = \Sigma x^2 - (\Sigma xy)^2 / \Sigma y^2 = 1,936 - (1,380)^2/1,000 = 31.60$ $s_y^2 = \Sigma d_y^2 / (n - 2) = 31.60/3 = 10.53$ $s_{y \cdot x} = \sqrt{10.53} = 3.245$ units $s_b = s_{y \cdot x} / \sqrt{\Sigma x^2} = 3.245/\sqrt{1,000} = 0.102$ $t = b/s_b = 1.38/0.102 = 13.5^{**}$, $df = n - 2 = 3$					

The figures shown under the sample data are all that need be written down. In most calculating machines, ΣX and ΣX^2 can be accumulated in a single run, ΣY and ΣY^2 in a second run and ΣXY in a third, without writing down any intermediate figures. With small samples in which X and Y have no more than three significant figures, some machines will accumulate ΣX , ΣY , ΣX^2 , $2\Sigma XY$, and ΣY^2 in one run.

EXAMPLE 6 3 1—The data show the initial weights and gains in weight (grams) of 15 female rats on a high protein diet from the 24th to 84th day of age. The point of interest in these data is whether the gain in weight depends to some extent on initial weight. If so, feeding experiments on female rats can be made more precise by taking account of the initial weights of the rats, either by pairing on initial weight or by adjusting for differences in initial weight in the analysis. Calculate b by the shortcut method and test its significance. Ans. $b = 1.0641$, $t = b/s_b = 2.02$ with 13 df , not quite significant at the 5% level.

	Rat Number														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Initial weight, X	50	64	76	64	74	60	69	68	56	48	57	59	46	45	65
Gain, Y	128	159	158	119	133	112	96	126	132	118	107	106	82	103	104

EXAMPLE 6.3.2—Speed records attained in the Indianapolis Memorial Day automobile races, 1911–1941, are as follows in miles per hour

Year	X	Speed Y	Year	X	Speed Y	Year	X	Speed Y
1911	0	74.6	1922	11	94.5	1932	21	104.1
1912	1	78.7	1923	12	91.0	1933	22	104.2
1913	2	75.9	1924	13	98.2	1934	23	104.9
1914	3	82.5	1925	14	101.1	1935	24	106.2
1915	4	89.8	1926	15	95.9	1936	25	109.1
1916	5	83.3	1927	16	97.5	1937	26	113.6
1917	6	*	1928	17	99.5	1938	27	117.2
1918	7	*	1929	18	97.6	1939	28	115.0
1919	8	88.1	1930	19	100.4	1940	29	114.3
1920	9	88.6	1931	20	96.6	1941	30	115.1
1921	10	89.6						

* No races

The years have been coded by subtracting 1911 from each. Calculate $\Sigma x^2 = 2,325.02$, $\Sigma y^2 = 4,039.81$, $\Sigma xy = 2,971.23$, $\bar{y} = 1.278\bar{x} + 77.57$ miles per hour

6.4—The mathematical model in linear regression. In standard linear regression, three assumptions are made about the relation between Y and X

- 1 For each selected X there is a normal distribution of Y from which the sample value of Y is drawn at random. If desired, more than one Y may be drawn from each distribution.
- 2 The population of values of Y corresponding to a selected X has a mean μ that lies on the straight line $\mu = \alpha + \beta(X - \bar{X}) = \alpha + \beta x$, where α and β are parameters (to be explained presently).
- 3 In each population the standard deviation of Y about its mean $\alpha + \beta x$ has the same value, often denoted by σ_y .

The mathematical model is specified concisely by the equation

$$Y = \alpha + \beta x + \varepsilon,$$

where ε is a random variable drawn from $N(0, \sigma_\varepsilon^2)$.

In this model, Y is the sum of a random part, ε , and a part fixed by x . The fixed part, according to assumption number 2 above, determines the means of the populations sampled, one mean for each x . These means lie on the straight line represented by $\mu = \alpha + \beta x$, the *population regression line*. The parameter α is the mean of the population that corresponds to $x = 0$, thus α specifies the height of the line when $x = \bar{X}$. β is the *slope* of the regression line, the *change* in Y per unit increase in x . As for the variable part of Y , ε is drawn at random from $N(0, \sigma_\varepsilon^2)$, it is independent of x and normally distributed, as the symbol N signifies.

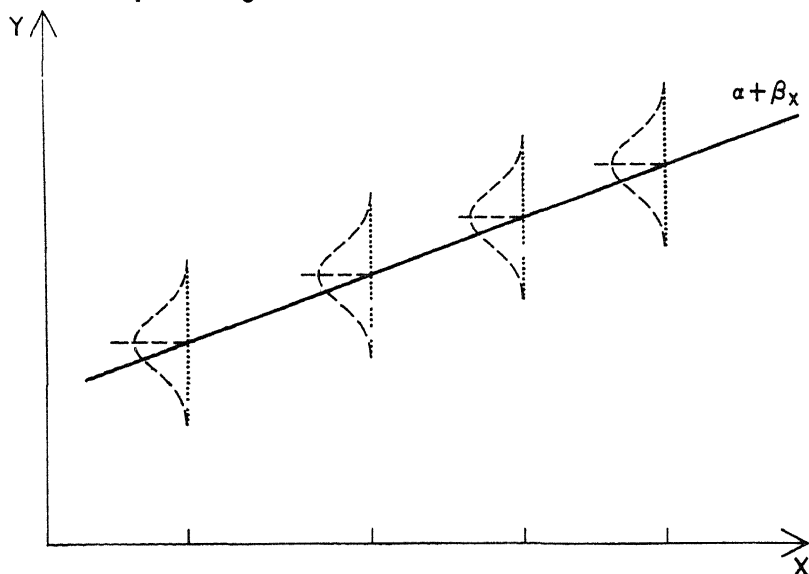


FIG 6.4.1—Representation of the linear regression model. The normal distribution of Y about the regression line $\alpha + \beta x$ is shown for four selected values of X .

Figure 6.4.1 gives a schematic representation of these populations. For each of four selected values of X the normal distribution of Y about its mean $\mu = \alpha + \beta x$ is sketched. These normal distributions would all coincide if their means were superimposed.

For non-mathematicians, the model is best explained by an arithmetical construction. Assign to X the values 0, 2, 3, 7, 8, 10, as in table 6.4.1. This is done quite arbitrarily; the manner in which X is fixed has no bearing on the illustration.

Next, calculate \bar{X} and the deviations, $x = X - \bar{X}$, in column 2.

Now take $\beta = 0.5$; this implies that the means of the populations are to increase one-half unit with each unit change in x . From this, column 3 is calculated.

Choose $\alpha = 4$, meaning that at $x = 0$ the population regression is 4 units above the X -axis.

The fixed X together with α and β determine the succession of means in column 4. These are indicated by open circles on the population regression line (the dotted line) of figure 6.4.2. So far all quantities are *fixed*, without sampling variation.

Coming now to the variable part of Y , the ε are drawn at random from a table of random normal deviates with mean zero $\sigma_{y \cdot x} = 1$. The values which we obtained were 1.1, -1.3, -1.1, 1.0, 0, and -1.0, as shown in column 5 of table 6.4.1. Column 6 contains the sample values of Y , each item being the sum of the fixed part in column 4 and the cor-

TABLE 6.4.1
CONSTRUCTION OF A SAMPLE FROM $Y = \alpha + \beta x + \varepsilon$, WITH $\alpha = 4$, $\beta = 0.5$,
AND ε DRAWN FROM $\mathcal{N}(0, 1)$

X	x	$\beta x = 0.5x$	$\alpha + \beta x = 4 + 0.5x$	ε	$Y = \alpha + \beta x + \varepsilon$
(1)	(2)	(3)	(4)	(5)	(6)
0	-5	-2.5	1.5	1.1	2.6
2	-3	-1.5	2.5	-1.3	1.2
3	-2	-1.0	3.0	-1.1	1.9
7	2	1.0	5.0	1.0	6.0
8	3	1.5	5.5	0.0	5.5
10	5	2.5	6.5	-1.0	5.5

Calculations of estimates for sample regression, Y on X

$\Sigma X' = 30$		$\Sigma Y = 22.7$
$\bar{X} = 5$		$\bar{Y} = 3.78$
$\Sigma \lambda^2 = 226$	$\Sigma XY = 149.1$	$\Sigma Y^2 = 108.31$
$(\Sigma X)^2/n = 150$	$\Sigma X \Sigma Y/n = 113.5$	$(\Sigma Y)^2/n = 85.88$
$\Sigma x^2 = 76$	$\Sigma xy = 35.6$	$\Sigma y^2 = 22.43$

$b = \Sigma xy / \Sigma x^2 = 35.6/76 = 0.468$	
$\hat{Y} = 3.78 + 0.468(X - 5) = 1.44 + 0.468X$	
$\Sigma d_{y,x}^2 = \Sigma y^2 - (\Sigma xy)^2 / \Sigma x^2 = 22.43 - (35.6)^2/76 = 5.75$	
$s_{y,x}^2 = \Sigma d_{y,x}^2 / (n - 2) = 5.75/4 = 1.44$	$s_{y,x} = \sqrt{1.44} = 1.20$

responding random part in column 5. The sample points are plotted in black circles in the figure.

The calculations of \bar{Y} and b are given under table 6.4.1. The popula-

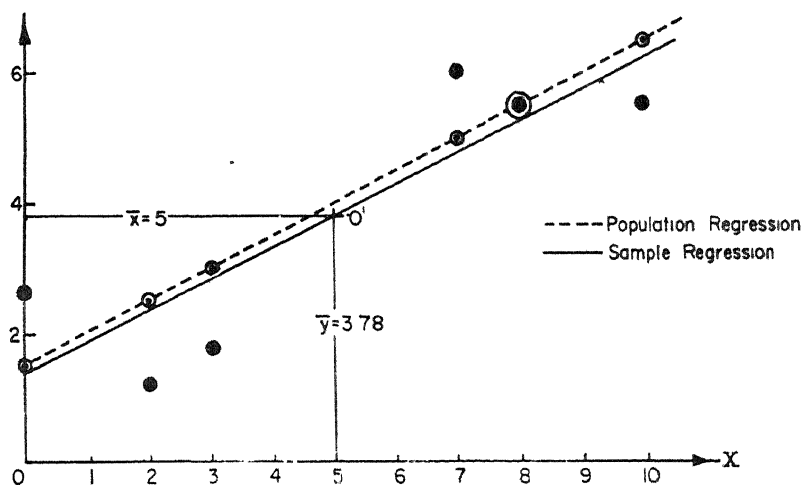


FIG 6.4.2 -Population regression $\mu = 4 + 0.5x$ Sample regression $\hat{Y} = 3.78 + 0.468x$

tion value $\alpha = 4$ is estimated by $\bar{Y} = 3.78$. The sample regression line passes through the point (\bar{X}, \bar{Y}) , (5, 3.78). The slope $\beta = 0.5$ is estimated by $b = 0.468$. The solid line in figure 6.4.2 is the sample regression line. It is nearly parallel to the population line but lies below it because of the underestimation of α . The discrepancies between the two lines are due wholly to the random sampling of the ε .

EXAMPLE 6.4.1—In table 6.4.1, $b = 0.468$. Calculate the six deviations from regression, $d_{y,x}$, and identify each with the distance of the corresponding point from the sample regression line. The sum of the deviations should be zero and the sum of their squares about 5.75.

EXAMPLE 6.4.2—Construct a sample with $\alpha = 6$ and $\beta = -1$. The negative β means that the regression will slope downwards to the right. Take $X = 1, 2, \dots, 9$, \bar{X} being 5. By using table 3.2.1, draw ε randomly from $\mathcal{N}(0, 5)$. Make a table showing the calculation of the sample of Y . Graph the population regression and the sample points. Save your work for further use.

6.5— \hat{Y} as an estimator of $\mu = \alpha + \beta x$. For any x , the computed value \hat{Y} estimates the corresponding $\mu = \alpha + \beta x$. For example, we have already seen that at $x = 0$ (for which $X = 5$), $\hat{Y}_5 = \bar{Y}$ estimates $\mu_5 = \alpha$. As another example, at $x = 2$, for which $X = 7$, $\hat{Y}_7 = 1.44 + (0.468)(7) = 4.72$, estimates $\mu = 4 + (0.5)(2) = 5$.

More generally,

$$\hat{Y} - \mu = (\bar{Y} - \alpha) + (b - \beta)x \quad (6.5.1)$$

Thus, the difference between \hat{Y} and the corresponding μ has two sources, both due to the random ε . One is the difference between the elevations of the sample and population regression lines $(\bar{Y} - \alpha)$; the other, the difference between the two slopes $(b - \beta)$.

Estimates of μ are often made at an X lying between two of the fixed X whose Y were sampled. For example, at $X = 4$,

$$\hat{Y}_4 = 1.44 + (0.468)(4) = 3.31,$$

locating a point on the sample regression line perpendicularly above $X = 4$. Here we are estimating μ in a population not sampled. There is no sample evidence for such an estimate; it is made on the cognizance of the investigator who has reason to believe that the intermediate population has a μ lying on the sampled regression, $\alpha + \beta x$.

Using the same argument, one may estimate μ at an X extrapolated beyond the range of the fixed X . Thus, at $X = 12$,

$$\hat{Y}_{12} = 1.44 + (0.468)(12) = 7.06$$

Extrapolation involves two extra hazards. Since x tends to be large for extrapolated values, equation 6.5.1 shows that the term $(b - \beta)x$ may make the difference $(\hat{Y} - \mu)$ large. Secondly (and this is usually the more serious hazard), the population regression of means may actually be curved to an extent that is small within the limits of the sample, but be-

comes pronounced when we move beyond these limits, so that results given by a straight-line regression are badly wrong.

The value of \hat{Y} also enables us to judge whether an individual observed Y is above or below its average value for the X in question. Look, for example, at the first point on the left of the graph (figure 6.4.2). $Y_0 = 2.6$, to be compared with $\hat{Y}_0 = 1.44$. The positive deviation, $d_{y \cdot 0} = Y_0 - \hat{Y}_0 = 1.16$, shows that Y_0 exceeds its estimated value by 1.16 units. Algebraically,

$$\begin{aligned} d_{y \cdot x} &= Y - \hat{Y} = \alpha + \beta x + \varepsilon - (\bar{Y} + bx) \\ &= (\alpha - \bar{Y}) + (\beta - b)x + \varepsilon \end{aligned}$$

Thus, $Y - \hat{Y}$ is, as would be expected, an estimate of the corresponding normal deviate ε , but is affected also by the errors in \bar{Y} and b . In the constructed example, $\varepsilon_0 = 1.1$, so that for this point $Y_0 - \hat{Y}_0 = 1.16$ is close. In large samples, the errors in \bar{Y} and b become small, and the residual $Y - \hat{Y}$ is a good estimate of the corresponding ε .

This examination of deviations from a fitted regression is often useful. A doctor's statement: "For a woman of your age, your blood pressure is normal," would imply that $Y - \hat{Y}$ was zero, or near to it. A value of Y that was quite usual in a woman aged 65 might cause a doctor to prescribe treatment if it occurred in a woman aged 35, because for this woman $Y - \hat{Y}$ would be exceptionally high.

EXAMPLE 6.5.1—For your sample in example 6.4.2, calculate \bar{Y} and b , then plot the sample regression line on your graph. Calculate the deviations $d_{y \cdot x}$ and compare them with the corresponding ε . It is a partial check on your accuracy if $\sum d_{y \cdot x} = 0$.

EXAMPLE 6.5.2—Using the blood pressure data of section 6.2, estimate μ at age 30 years. Ans. 106.5 units.

EXAMPLE 6.5.3—Calculate $Y_A = Y - bx$, called *adjusted Y*, for each age group in table 6.2.2. Verify your results by the sum, $\sum Y_A = \sum Y$. Suggest several possible reasons for the differences among adjusted Y .

6.6—The estimator of $\sigma_{y \cdot x}^2$. As noted earlier, the quantity

$$s_{y \cdot x}^2 = \sum d_{y \cdot x}^2 / (n - 2)$$

is an unbiased estimator of $\sigma_{y \cdot x}^2$, the variance of the ε 's. One way of remembering the divisor $(n - 2)$ is to note that in fitting the line we have two disposable constants, α and β , whose values we choose so to make the $d_{y \cdot x}$ as small as possible. If there are only two points (Y_1, X_1) and (Y_2, X_2), the fitted line goes through both points *exactly*. The $d_{y \cdot x}$ and their sum of squares are then zero, no matter how large the true $\sigma_{y \cdot x}$ is. In other words, there are no degrees of freedom remaining for estimating $\sigma_{y \cdot x}^2$.

In the constructed example (table 6.4.1), $s_{y \cdot x}^2$ was found to be 1.44, with 4 d.f., as an estimate of $\sigma_{y \cdot x}^2 = 1$. This gives 1.20 as the estimate of $\sigma_{y \cdot x} = 1$.

The estimated variance in the original sample of values of Y is $s_y^2 = 22.43/5 = 4.49$. By utilization of the knowledge of X , this variance

is reduced to $s_{y \cdot x}^2 = 1.44$. It is sometimes said that a fraction $(4.49 - 1.4)/4.49$, or about 68% of the variation in Y is associated with the linear regression on X , the remaining 32% being independent of X . This statement is useful when the objective is to understand why Y varies and it is known that X is one of the causes of the variation in Y .

The nature of $s_{y \cdot x}^2$ is also made clearer by some algebra. For the i th member of the sample,

$$\varepsilon_i = Y_i - \alpha - \beta x_i \quad ; \quad d_{y \cdot x_i} = Y_i - \bar{Y} - bx_i = y_i - bx_i$$

Write

$$\begin{aligned} \varepsilon_i &= Y_i - \alpha - \beta x_i = Y_i - \bar{Y} - bx_i + (\bar{Y} - \alpha) + (b - \beta)x_i \\ &= (y_i - bx_i) + (\bar{Y} - \alpha) + (b - \beta)x_i \end{aligned}$$

Square both sides and sum over the n values in the sample. On the right side there are three squared terms and three product terms. The squared terms give

$$\Sigma(y_i - bx_i)^2 + \Sigma(\bar{Y} - \alpha)^2 + \Sigma(b - \beta)^2 x_i^2$$

The factors $(\bar{Y} - \alpha)^2$ and $(b - \beta)^2$ are constant for all members of the sample and can be taken outside the Σ sign. This gives, for the squared terms,

$$\Sigma(y_i - bx_i)^2 + n(\bar{Y} - \alpha)^2 + (b - \beta)^2 \Sigma x_i^2$$

Remarkably, the three cross-product terms all vanish when summed over the sample. For example,

$$2\Sigma(y_i - bx_i)(\bar{Y} - \alpha) = 2(\bar{Y} - \alpha)\Sigma(y_i - bx_i) = 0$$

since $\Sigma y_i = 0$ and $\Sigma x_i = 0$. Further,

$$2\Sigma(\bar{Y} - \alpha)(b - \beta)x_i = 2(\bar{Y} - \alpha)(b - \beta)\Sigma x_i = 0,$$

$$\begin{aligned} 2\Sigma(y_i - bx_i)(b - \beta)x_i &= 2(b - \beta)\Sigma x_i(y_i - bx_i) \\ &= 2(b - \beta)(\Sigma x_i y_i - b \Sigma x_i^2) \end{aligned}$$

which vanishes since $b = \Sigma x_i y_i / \Sigma x_i^2$. Thus, finally,

$$\begin{aligned} \Sigma \varepsilon_i^2 &= \Sigma(Y_i - \alpha - \beta x_i)^2 = \Sigma(Y_i - \bar{Y} - bx_i)^2 + n(\bar{Y} - \alpha)^2 \\ &\quad + (b - \beta)^2 \Sigma x_i^2 \end{aligned} \quad (6.6.1)$$

Rearranging,

$$\Sigma d_{y \cdot x}^2 = \Sigma(Y_i - \bar{Y} - bx_i)^2 = \Sigma \varepsilon_i^2 - n(\bar{Y} - \alpha)^2 - (b - \beta)^2 \Sigma x_i^2$$

On the right side of this equation, each ε_i has mean zero and variance $\sigma_{y \cdot x}^2$. Thus the term $\Sigma \varepsilon_i^2$ is an estimate of $n\sigma_{y \cdot x}^2$. The two subtracted terms on the right can be shown to be estimates of $\sigma_{y \cdot x}^2$. It follows that $\Sigma d_{y \cdot x}^2$ is an unbiased estimate of $(n - 2)\sigma_{y \cdot x}^2$, and on division by $(n - 2)$ provides an unbiased estimate of $\sigma_{y \cdot x}^2$. This result, namely that $s_{y \cdot x}^2$ is unbiased, does not require the ε_i to be normally distributed. Normality is required, however, to prove the standard tests of significance in regression.

6.7—The method of least squares. The choice of \bar{Y} and b to estimate the parameters α and β is an application of a principle widely used in problems of statistical estimation and known as *the method of least squares*. To explain this method, let $\hat{\alpha}$ and $\hat{\beta}$ denote any two estimators of α and β that we might consider. For the pair of observations (Y, X) the quantity

$$Y - \hat{\alpha} - \hat{\beta}x$$

measures the amount by which the fitted regression is in error in estimating Y . In the method of least squares, $\hat{\alpha}$ and $\hat{\beta}$ are chosen so as to minimize the sum of the squares of these errors, taken over the sample. That is, we minimize

$$\Sigma(Y - \hat{\alpha} - \hat{\beta}x)^2 \quad (6.7.1)$$

About 150 years ago the scientist Gauss showed that estimators obtained in this way are (i) unbiased, and (ii) have the smallest standard errors of any unbiased estimators that are linear expressions in the Y 's. Gauss' proof does not require the Y 's to be normally distributed, but merely that the ε 's are independent with means zero and variances $\sigma_{y \cdot x}^2$.

The result that (6.7.1) is minimized by taking $\hat{\alpha} = \bar{Y}$ and $\hat{\beta} = b$ is easily verified by quoting a previous result (6.6.1, p. 146). Since the proof of the algebraic equality in (6.6.1) may be shown to hold for any pair of values α, β , the equation remains valid if we replace α by $\hat{\alpha}$ and β by $\hat{\beta}$. Hence quoting (6.6.1),

$$\Sigma(Y - \hat{\alpha} - \hat{\beta}x)^2 = \Sigma(Y - \bar{Y} - bx)^2 + n(\bar{Y} - \hat{\alpha})^2 + (b - \hat{\beta})^2 \Sigma x^2$$

The first term on the right is the sum of squares of the errors or residuals that we obtain if we take $\hat{\alpha} = \bar{Y}$ and $\hat{\beta} = b$. The two remaining terms on the right are both positive unless $\hat{\alpha} = \bar{Y}$ and $\hat{\beta} = b$. This proves that the choice of \bar{Y} and b minimizes (6.7.1).

6.8—The value of b in some simple cases. The expression for b , $\Sigma xy / \Sigma x^2$ is unfamiliar at first sight. It is not obviously related to the quantity β of which b is an estimate, nor is it clear that this is the estimate that common sense would suggest to someone who had never heard of least squares. A general expression relating b and β and an examination of a few simple cases may make b more familiar.

Denote the members of the sample by (Y_i, X_i) , where the subscript i goes from 1 to n . The numerator of b is $\Sigma x_i y_i = \Sigma x_i (Y_i - \bar{Y}) = \Sigma x_i Y_i - \Sigma x_i \bar{Y}$. Since the term $\Sigma x_i \bar{Y}$ vanishes, because $\Sigma x_i = 0$, the numerator of b may be written $\Sigma x_i Y_i$. Now substitute $Y_i = \alpha + \beta x_i + \varepsilon_i$. This gives

$$b = \frac{\Sigma x_i (\alpha + \beta x_i + \varepsilon_i)}{\Sigma x_i^2} = \beta \frac{\Sigma x_i^2}{\Sigma x_i^2} + \frac{\Sigma x_i \varepsilon_i}{\Sigma x_i^2} = \beta + \frac{\Sigma x_i \varepsilon_i}{\Sigma x_i^2},$$

the term in α vanishing because $\Sigma x_i = 0$. Thus b differs from β by a linear

expression in the ε_i . If the ε_i were all zero, b would coincide with β . Further, since the ε_i have zero means in the population, it follows that b is an unbiased estimate of β .

Turning to the simplest case, suppose that the sample consists of the values $(Y_1, 1)$ and $(Y_2, 2)$. The obvious estimate of the change in Y per unit increase in X is $Y_2 - Y_1$. What does b give? Since $\bar{X} = 1\frac{1}{2}$, the deviations are $x_1 = -1/2$, $x_2 = +1/2$, giving $\sum x^2 = 1/2$. Thus

$$b = \frac{-\frac{1}{2} Y_1 + \frac{1}{2} Y_2}{\frac{1}{2}} = Y_2 - Y_1,$$

in agreement.

With three values $(Y_1, 1)$, $(Y_2, 2)$, $(Y_3, 3)$ we might argue that $Y_2 - Y_1$ and $Y_3 - Y_2$ are both estimates of the change in Y per unit change in X . Since there seems no reason to do otherwise, we might average them, getting $(Y_3 - Y_1)/2$ as our estimate. To compare this with the least squares estimate, we have $x_1 = -1$, $x_2 = 0$, $x_3 = +1$. This gives $\sum xY = Y_3 - Y_1$ and $\sum x^2 = 2$, so that $b = (Y_3 - Y_1)/2$, again in agreement with the common-sense approach. Notice that Y_2 is not used in estimating the slope. Y_2 is useful in providing a check on whether the population regression line is straight. If it is straight, Y_2 should be equal to the average of Y_1 and Y_3 , apart from sampling errors. The difference $Y_2 - (Y_1 + Y_3)/2$ is therefore a measure of the curvature (if any) of the population regression.

Continuing in this way for the sample $(Y_1, 1)$, $(Y_2, 2)$, $(Y_3, 3)$, $(Y_4, 4)$, we have three simple estimates of β , namely $(Y_2 - Y_1)$, $(Y_3 - Y_2)$, and $(Y_4 - Y_3)$. If we average them as before, we get $(Y_4 - Y_1)/3$. This is disconcerting, since this estimate does not use either Y_2 or Y_3 . What does least squares give? The values of x are $-3/2$, $-1/2$, $+1/2$, and $+3/2$ and the estimate may be verified to be

$$b = (3Y_4 + Y_3 - Y_2 - 3Y_1)/10.$$

The least squares result can be explained as follows. The quantity $(Y_4 - Y_1)/3$ is an estimate of β , with variance $2\sigma_{y,x}^2/9$. The sample supplies another independent estimate $(Y_3 - Y_2)$, with variance $2\sigma_{y,x}^2$. In combining these two estimates, the principle of least squares weights them inversely as their variances, assigning greater weight to the more accurate estimate. This weighted estimate is

$$[9(Y_4 - Y_1)/3 + (Y_3 - Y_2)]/(9 + 1) = (3Y_4 + Y_3 - Y_2 - 3Y_1)/10 = b$$

As these examples show, it is easy to construct unbiased estimates of β by simple, direct methods. The least squares approach automatically produces the estimate with the smallest standard error.

Remember that b estimates the *average change in Y per unit increase in X* . Reporting a value of b requires that both units be stated, such as "systolic blood pressure per year of age."

6.9—The situation when X varies from sample to sample. Often the investigator does not select the values of X . Instead, he draws a sample from some population, then measures two characters Y and X for each member of the sample. In our illustration, the sample is a sample of apple trees in which the relation between the percentage of wormy fruits Y on a tree and the size X of its fruit crop is being investigated. In such applications the investigator realizes that if he drew a second sample, the values of X in that sample would differ from those in the first sample. In the results presented in preceding sections, we regarded the values of X as essentially fixed. The question is sometimes asked: can these results be used when it is known that the X -values will change from sample to sample?

Fortunately, the answer is yes, provided that for any value of X the corresponding Y satisfies the three assumptions stated at the beginning of section 6.4. For each X , the sample value of Y must be drawn from a normal population that has mean $\mu = \alpha + \beta x$ and constant variance $\sigma_{y \cdot x}^2$. Under these conditions the calculations for fitting the line, the t -test of b , and the methods given later to construct confidence limits for β and for the position of the true line all apply without change.

Consider, for instance, the accuracy with which β is estimated by b . The standard error of b is $\sigma_{y \cdot x}/\sqrt{(\Sigma x^2)}$. If a second sample of n apple trees were to be drawn, we know that Σx^2 , and hence the standard error of b , would change. That is, when X varies from sample to sample, some samples of size n provide more accurate estimates of β than others. But since the value of Σx^2 is known for the sample actually drawn, it makes sense to attach to b the standard error $\sigma_{y \cdot x}/\sqrt{(\Sigma x^2)}$, or its estimate $s_{y \cdot x}/\sqrt{(\Sigma x^2)}$. By doing so we take account of the fact that our b may be somewhat more accurate or somewhat less accurate than is usual in a sample of size n . In statistical theory this approach is sometimes described as using the *conditional* distribution of b for the values of X that we obtained in our sample, rather than the general distribution of b in repeated samples of size n .

There is one important distinction between the two cases. Suppose that in a study of families, the heights of pairs of adult brothers (X) and sisters (Y) are measured. An investigator might be interested either in the regression of sister's height on brother's height:

$$\hat{Y} = \bar{Y} + b_{y \cdot x}(X - \bar{X})$$

or in the regression of brother's height on sister's height:

$$\hat{X} = \bar{X} + b_{x \cdot y}(Y - \bar{Y})$$

These two regression lines are *different*. For a sample of 11 pairs of brothers and sisters, they are shown in figure 7.1.1 (p. 173). The line AB in this figure is the regression of Y on X , while the line CD is the regression of X on Y . Since $b_{y \cdot x} = \Sigma xy / \Sigma x^2$ and $b_{x \cdot y} = \Sigma xy / \Sigma y^2$, it follows that $b_{x \cdot y}$ is not in general equal to $1/b_{y \cdot x}$, as it would have to be to make the slopes AB and CD identical.

If the sample of pairs (X, Y) is a random one, the investigator may use whichever regression is relevant for his purpose. In predicting brother's heights from sister's heights, for instance, he uses the regression of X on Y . If, however, he has deliberately selected his sample of values of one of the variates, say X , then only the regression of Y on X has meaning and stability. There are many reasons for selecting the values of X . The levels of X may represent different amounts of a drug to be applied to groups of animals, or persons of ages 25, 30, 35, 40, 45, selected for convenience in calculating and graphing the regression of Y on age, or a deliberate choice of extremes, so as to make Σx^2 large and decrease the standard error of b , $\sigma_{y \cdot x} / \sqrt{(\Sigma x^2)}$. Provided that the X are selected without seeing the corresponding Y values, the linear regression line of Y on X is not distorted. Selection of the Y values, on the other hand, can greatly change this regression. Clearly, if we choose Y values that are all equal, the sample regression b of Y on X will be zero whatever the slope of the population regression.

To turn to the numerical example, it contains another feature of interest, a regression that is negative instead of positive.

TABLE 6.9.1
REGRESSION OF PERCENTAGE OF WORMY FRUIT ON SIZE OF APPLE CROP

Tree Number	Size of Crop on Tree (hundreds of fruits) X	Percentage of Wormy Fruits Y	Estimate of μ \hat{Y}	Deviation From Regression $Y - \hat{Y} = d_{y \cdot x}$
1	8	59	56.14	2.86
2	6	58	58.17	-0.17
3	11	56	53.10	2.90
4	22	53	41.96	11.04
5	14	50	50.06	-0.06
6	17	45	47.03	-2.03
7	18	43	46.01	-3.01
8	24	42	39.94	2.06
9	19	39	45.00	-6.00
10	23	38	40.95	-2.95
11	26	30	37.91	-7.91
12	40	27	23.73	3.27

$$\begin{array}{lll}
 \Sigma X = 228 & \Sigma Y = 540 & \\
 \bar{X} = 19 & \bar{Y} = 45 & \\
 \Sigma X^2 = 5,256 & \Sigma Y^2 = 25,522 & \Sigma XY = 9,324 \\
 (\Sigma X)^2/n = 4,332 & (\Sigma Y)^2/n = 24,300 & (\Sigma X)(\Sigma Y)/n = 10,260
 \end{array}$$

$$\begin{array}{lll}
 \Sigma x^2 = 924 & \Sigma y^2 = 1,222 & \Sigma xy = -936 \\
 b = \Sigma xy / \Sigma x^2 = -936/924 = -1.013 \text{ per cent per 100 wormy fruits} & & \\
 \hat{Y} = \bar{Y} + b(X - \bar{X}) = 45 - 1.013(X - 19) = 64.247 - 1.013X & & \\
 \Sigma d_{y \cdot x}^2 = 1,222 - (-936)^2/924 = 273.88 & & \\
 s_{y \cdot x}^2 = \Sigma d_{y \cdot x}^2 / (n - 2) = 273.88/10 = 27.388 & &
 \end{array}$$

It is generally thought that the percentage of fruits attacked by codling moth larvae is greater on apple trees bearing a small crop. Apparently the density of the flying moth tends towards uniformity, so that the chance of attack for any particular fruit is augmented if there are few fruits in the tree. The data in table 6.9.1 are adapted from the results of an experiment (3) containing evidence about this phenomenon. The 12 trees were all given a calyx spray of lead arsenate followed by 5 cover sprays made up of 3 pounds of manganese arsenate and 1 quart of fish oil per 100 gallons. There is a decided tendency, emphasized in figure 6.9.1, for the percentage of wormy fruits to decrease as the number of apples in the tree increases. In this particular group of trees, the relation of the two variates is even closer than usual.

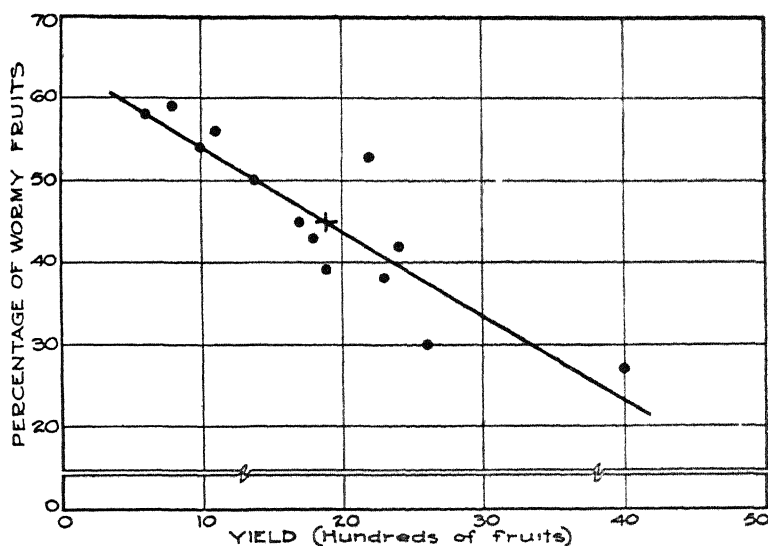


FIG. 6.9.1 -Sample regression of percentage of wormy fruits on size of crop in apple trees. The cross indicates the origin for deviations, $0'(\bar{X}, \bar{Y})$

The new feature in the calculations is the majority of negative products, xy , caused by the tendency of small values of Y to be associated with large values of X . The sample regression coefficient shows that the estimated percentage of wormy apples decreases, as indicated by the minus sign, 1.013 with each increase of 100 fruits in the crop. The sample regression line, and of course the percentage, falls away from the point, $0'(\bar{X}, \bar{Y})$, by 1.013 for each unit of crop above 19 hundreds.

The regression line brings into prominence the deviations from this moving average, deviations which measure the failure of crop size to account for variation in the intensity of infestation. Trees number 4, 9, and 11 had notably discrepant percentages of injured fruits, while numbers 2

and 5 performed as expected. According to the model these are random deviations from the average (regression) values, but close observation of the trees during the flight of the moths might reveal some characteristics of this phenomenon. Tree 4 might have been on the side from which the flight originated or perhaps its shape or situation caused poor applications of the spray. Trees 9 and 11 might have had some peculiarities of conformation of foliage that protected them. Careful study of trees 2 and 5 might throw light on the kind of tree or location that receives normal infestation. This kind of *case study* usually does not affect the handling of the sample statistics, but it may add to the investigator's knowledge of his experimental material and may afford clues to the improvement of future experiments.

Among attitudes toward experimental data, two extremes exist, both of which should be avoided: some attend only to minute details of sample variation, neglecting the summarization of the data and the consequent inferences about the population; others are impatient of the data themselves, rushing headlong toward averages and other generalizations. Either course fails to yield full information from the experiment. The competent investigator takes time to examine each datum together with the individual measured. He attempts to distinguish normal variation from aberrant observations. He then appraises his summary statistics and his population inferences and draws his conclusions against this background of sample facts.

EXAMPLE 6.9.1—Another group of 12 trees, investigated by Hansberry and Richardson, was sprayed with lead arsenate throughout the season. In addition, the fourth and fifth cover sprays contained 1% mineral oil emulsion and nicotine sulfate at the rate of 1 pint per 100 gallons. The results are shown below. These facts may be verified: $\Sigma X = 240$, $\Sigma Y = 384$, $\Sigma x^2 = 808$, $\Sigma y^2 = 1,428$, $\Sigma xy = -582$, regression coefficient = -0.7203 , $\hat{Y} = 46.41 - 0.7203X$, $Y - \hat{Y}$ for the first tree = 16.40%.

Size of Crop, X Hundreds	15, 15, 12, 26, 18, 12, 8, 38, 26, 19, 29, 22
Percentage Wormy, Y	52, 46, 38, 37, 37, 37, 34, 25, 22, 22, 20, 14

EXAMPLE 6.9.2—In table 6.9.1, calculate $\Sigma d_{y,x}^2 = 273.88$ by means of the formula given in section 6.2.

EXAMPLE 6.9.3—The following weights of body and comb of 15-day-old White Leghorn male chicks are adapted from Snedecor and Breneman (4).

Chick Number	1	2	3	4	5	6	7	8	9	10
Body weight (grams), X	83	72	69	90	90	95	95	91	75	70
Comb weight (milligrams), Y	56	42	18	84	56	107	90	68	31	48

Calculate the sample regression equation, $\hat{Y} = 60 + 2.302(X - 83)$.

EXAMPLE 6.9.4—Construct the graph of the chick data, plotting body weight along the horizontal axis. Insert the regression line.

6.10—Interval estimates of β and tests of null hypotheses. Being provided with point estimates of the parameters of the regression population, we turn to their interval estimates and to tests of hypotheses about them.

First in order of utility, there is the sample regression coefficient b , an estimate of β . As seen in section 6.2, in random sampling, b is distributed with a variance estimated by

$$s_b^2 = s_{y \cdot x}^2 / \Sigma x^2$$

Thus, in the apple sampling of table 6.9.1,

$$s_b^2 = 27.388/924 = 0.0296; s_b = 0.172\%$$

Moreover, since the quantity $(b - \beta)/s_b$ follows the t -distribution with $n - 2$ degrees of freedom, it may be said with 95% confidence that

$$b - t_{0.05} s_b \leq \beta \leq b + t_{0.05} s_b$$

For the apples, $d.f. = 10$, $t_{0.05} = 2.228$, $t_{0.05} s_b = (2.228)(0.172) = 0.383$,

$$b - t_{0.05} s_b = -1.013 - 0.383 = -1.396 \text{ per cent per 100 fruits,}$$

$$b + t_{0.05} s_b = -1.013 + 0.383 = -0.630 \text{ per cent per 100 fruits,}$$

and, finally,

$$-1.396 \leq \beta \leq -0.630$$

If it is said that the population regression coefficient is within these limits, the statement is right unless the sample is one of the divergent kind that occurs about once in 20 trials.

Instead of the interval estimate of β , interest may lie in testing some null hypothesis. While it is now rather obvious that $H_0: \beta = 0$ will be rejected, we proceed with the illustration; if there were any other pertinent value of β to be tested, we could use that instead. Since $(b - \beta)/s_b$ follows the t -distribution we put

$$t = \frac{b - \beta}{s_b} = \frac{-1.013 - 0}{0.172} = -5.89, \quad d.f. = n - 2 = 10$$

The sign is ignored because the table contains both halves of the distribution. H_0 is rejected. One concludes that in the population sampled there is a regression of percentage wormy apples on crop size, the value likely being between -0.630 and -1.396 per cent per 100 fruits.

6.11—Prediction of the population regression line. Next, we may wish to make inferences about $\mu = \alpha + \beta x$, that is, about the height of the population regression line at the point X . The sample estimate of μ is $\hat{Y} = \bar{Y} + bx$. The error in the prediction is

$$\hat{Y} - \mu = (\bar{Y} - \alpha) + (b - \beta)x$$

But since $Y = \alpha + \beta x + \varepsilon$, we have $\bar{Y} = \alpha + \bar{\varepsilon}$, giving

$$\hat{Y} - \mu = \bar{\varepsilon} + (b - \beta)x \quad (6.11.1)$$

The term $\bar{\varepsilon}$ has variance $\sigma_{y \cdot x}^2/n$. Further, b is distributed about β with variance $\sigma_{y \cdot x}^2/\Sigma x^2$. Finally, the independence of the ε 's guarantees that these two sources of error are uncorrelated, so that the variance of their sum is the sum of the two variances. This gives

$$\sigma_{\hat{Y}}^2 = \sigma_{y \cdot x}^2 \left(\frac{1}{n} + \frac{x^2}{\Sigma x^2} \right)$$

The estimated standard error of \hat{Y} is

$$s_{\hat{Y}} = s_{y \cdot x} \sqrt{(1/n) + (x^2/\Sigma x^2)} \quad (6.11.2)$$

with $(n - 2)$ d.f.

For the apples, $s_{y \cdot x} = \sqrt{27.388}$, $n = 12$, and $\Sigma x^2 = 924$.

$$s_{\hat{Y}} = \sqrt{27.388} \sqrt{(1/12) + (x^2/924)} = \sqrt{2.282 + 0.02964x^2}$$

For trees with a high crop like that of Tree 12, $x = 21$ and $s_{\hat{Y}} = 3.92\%$, notably greater than $s_{\hat{Y}} = 1.51\%$ at $x = 0$. The reason why $s_{\hat{Y}}$ increases as X recedes from \bar{X} is evident from the term $(b - \beta)x$ in equation (6.11.1). The effect of any error in b is steadily magnified as x becomes greater.

Corresponding to any \hat{Y} , the point estimate of μ , there is an interval estimate

$$\hat{Y} - t_{0.05} s_{\hat{Y}} \leq \mu \leq \hat{Y} + t_{0.05} s_{\hat{Y}}$$

One might wish to estimate the mean percentage of wormy apples, μ , at the point $X = 30$ hundreds of fruits. If so,

$$x = X - \bar{X} = 30 - 19 = 11 \text{ hundreds of fruits}$$

$$\hat{Y} = \bar{Y} + bx = 45 - (1.013)(11) = 33.86\%$$

$$t_{0.05} s_{\hat{Y}} = (2.228) \sqrt{2.282 + (0.02964)(11^2)} = 5.40\%$$

$$33.86 - 5.40 \leq \mu \leq 33.86 + 5.40$$

Finally,

$$28.46\% \leq \mu \leq 39.26\%$$

At $X = 30$ hundreds of fruits, the population mean μ is estimated as 33.86% wormy fruits with 0.95 confidence limits from 28.46% to 39.26%. This confidence interval is represented by AB in figure 6.11.1.

If calculations like this are done for various values of X and if the confidence limits are plotted above and below the sample regression line, one has a confidence belt or zone with curved borders DB and CA in figure 6.11.1. The curves are the branches of a hyperbola. We have

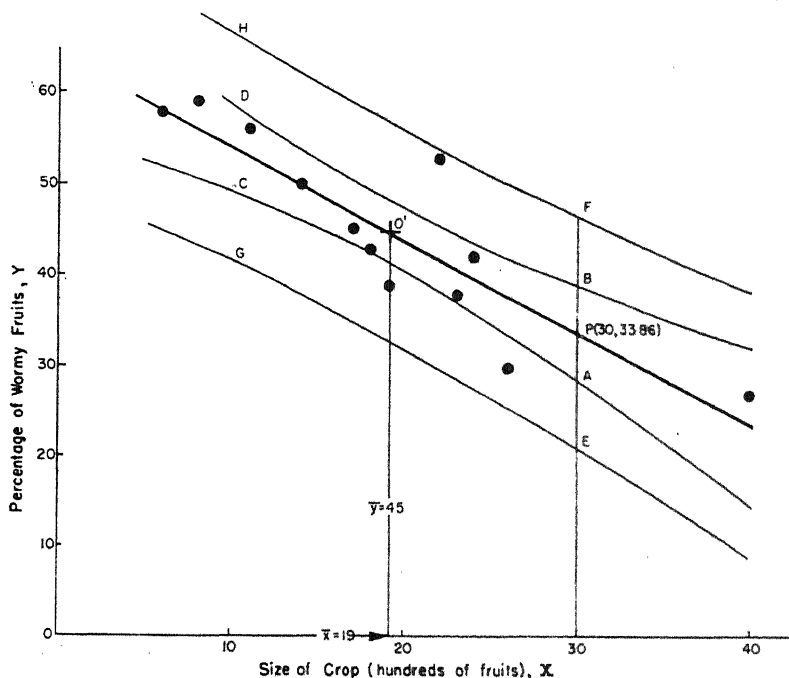


FIG. 6.11.1—Confidence belts for μ , $ABCD$; and for Y , $EFGH$; the apple data.

confidence that μ_y for any X lies in the belt. The figure emphasizes the increasing hazard of making predictions at X far removed from \bar{X} .

6.12—Prediction of an individual Y . A further use of regression is to predict the individual value of Y for a new member of the population for which X has been measured. The predicted value is again $\hat{Y} = \bar{Y} + bx$, but since $Y = \alpha + \beta x + \varepsilon$, the error of the prediction now becomes

$$\hat{Y} - Y = (\bar{Y} - \alpha) + (b - \beta)x - \varepsilon$$

The random element ε for the new member is an additional source of uncertainty. So, the mean square error of the predicted value contains another term, being

$$s_{\hat{Y}}^2 = \frac{s_{y \cdot x}^2}{n} + \frac{x^2 s_{y \cdot x}^2}{\sum x^2} + s_{y \cdot x}^2$$

Since the term arising from the variance of ε usually dominates, the standard error is usually written as

$$s_{\hat{Y}} = s_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{x^2}{\sum x^2}} \quad (6.12.1)$$

It is important not to confuse the two types of prediction. If the regression of weight on height were worked out for a sample of 20-year-old males, the purpose might be to predict the *average* weight of 20-year-old males of a specific height. This is prediction of μ given X . Alternatively, we might want to predict the weight of a new male whose height is known. This is prediction of an individual Y , given X .

The two prediction problems have the interesting feature that the prediction, \hat{Y} , is exactly the same in the two problems, but the standard error of the prediction differs (compare equations [6.11.2] and [6.12.1]). To avoid confusion, use the symbols $\hat{\mu}$ and $s_{\hat{\mu}}$ when a population average is being predicted, and \hat{Y} and $s_{\hat{Y}}$ when an individual Y is being predicted. For example, if you wish to predict the percentage of wormy apples on a tree yielding 30 hundreds of fruits,

$$t_{0.05s_{\hat{Y}}} = 2.228 \sqrt{27.388} \sqrt{1 + 1/12 + (11)^2/924} = 12.85\%$$

From this and $\hat{Y} = 33.86\%$, the confidence interval is given by

$$33.86 - 12.85 \leq Y \leq 33.86 + 12.85$$

or,

$$21.01\% \leq Y \leq 46.71\%,$$

as shown by *EF*, figure 6.11.1. We conclude that for trees bearing 3,000 fruits, population values of percentage wormy fruits fall between 21.01% and 46.71% unless a 1-in-20 chance has occurred in the sampling.

Continuing this procedure, a confidence belt *HF* and *GE* for Y may be plotted as in the figure. It is to be observed that all the sample points lie in the belt. In general about 5% of them are expected to fall outside.

Unfortunately, the meaning of this confidence band is apt to be misunderstood. Suppose that we construct 95% confidence intervals for the Y values of a large number of new individual specimens that all have the same value of X . The 95% confidence probability is correct if for each new specimen we draw a *new* sample of values of (Y, X) , compute a *new* sample regression line and value of $s_{y \cdot x}$, and construct a *new* confidence interval from these data. If we make a large number of confidence interval statements from the *same* sample regression line, the proportion of these statements that is correct is not 95% for a specific line, but may be more or less. If the sample from which the regression line was computed happens to give an unusually low value of $s_{y \cdot x}$, so that the confidence band is narrower than usual, less than 95% of the confidence interval statements is likely to be correct.

This point can be illustrated from the line constructed in table 6.4.1 (p. 143) as an example of the regression model. The sample line is $1.44 + 0.468X$, and has a value 2.376 at $X = 2$. Further, $s_{\hat{Y}}$ at $X = 2$ is found to be 1.325, and $t_{0.05}$, for 4 *d.f.*, is 2.776. Hence, the 95% confidence limits for an individual Y at $X = 2$ are $2.376 \pm (2.776)(1.325)$, giving -1.302 and 6.054

But we know from the population model that any new Y at $X = 2$ is normally distributed with $\mu = 5$ and $\sigma = 1$. The probability that this Y lies between 0.948 and 8.484 is easily calculated from the normal table. It is practically 100%, instead of 95%. In fact, with this sample line, the 95% confidence probability statements are conservative in this way at all six values of X .

The worker who makes many predictions from the same sample line naturally wants some kind of probability statement that applies to his line. The available techniques are described by Acton (11).

EXAMPLE 6.12.1—In the regression of comb weight of chicks on body weight, example 6.9.3, $n = 10$, $\bar{X} = 83$ gms., $\bar{Y} = 60$ mg., $\Sigma x^2 = 1,000$, $\Sigma y^2 = 6,854$ and $\Sigma xy = 2,302$. Set 95% confidence limits on α , assuming the same set of body weights. Ans. 49.8 – 70.2 mg

EXAMPLE 6.12.2—In the chick data, $b = 2.302$. Test the hypothesis that $\beta = 0$. Ans. $t = 5.22$, $P < 0.01$.

EXAMPLE 6.12.3—Since evidently there is a population regression of comb weight on body weight, set 95% limits to the regression coefficient. Ans. 1.28 – 3.32 mg. per gm

EXAMPLE 6.12.4—Predict the population average comb weight of 100-gm. chicks. Ans. 99.1 mg. with 95% limits, 79.0 – 119.2 mg.

EXAMPLE 6.12.5—Set 95% confidence limits to the forecast of the comb weight of a randomly chosen 100-gm. chick. Ans. 61.3 – 136.9 mg.

EXAMPLE 6.12.6—In the Indianapolis motor races (example 6.3.2) estimate the speed for the year 1946, for which the coded X is 35, and give 95% limits, remembering that individual speeds are being estimated. Ans. 122.3 miles per hour with 95% limits 118.9 – 125.7. The actual speed in 1946 was 114.8 miles per hour, lying outside the limits. The regression formula overestimated the speeds consistently in the ten years following 1945.

EXAMPLE 6.12.7—Construct 80% confidence bands for the individual race results in the period 1911–1941. Since there were 29 races, you should find about 6 results lying outside the band.

EXAMPLE 6.12.8—In time series such as these races, the assumption that the ϵ are independent of each other may not hold. Winning of successive races by the same man, type of car, or racing technique, all raise doubts on this point. If the ϵ are not independent, \bar{Y} and b remain unbiased estimates of α and β , but they are no longer the most precise estimates, and the formulas for standard errors and confidence limits become incorrect.

6.13—Testing a deviation that looks suspiciously large. When Y is plotted against X , one or two points sometimes look as if they lie far from the regression line. When the line has been computed, we can examine this question further by drawing the line and looking at the deviations for these points, or by calculating the values of $d_{y,x}$ for them.

In this process one needs some guidance with respect to the question: When is a deviation large enough to excite suspicion? A test of significance is carried out as follows:

1. Select the point with the largest $d_{y,x}$ (in absolute value). As an illustration, we use the regression of wormy fruit on size of apple crop, table 6.9.1 and figure 6.9.1, p. 151. We have already commented that for tree 4, with $X = 22$, $Y = 53$, the deviation $d_{y,x} = 11.04$ looks large.

2. Recompute the regression with this point omitted. This requires little work, since from the values ΣX , ΣY , ΣX^2 , ΣY^2 , and ΣXY , we simply subtract the contribution for tree 4. We find for the remaining $n - 1 = 11$ points:

$$\begin{aligned}\bar{X} &= 18.73 : \Sigma x^2 = 914 \\ \hat{Y} &= 44.27 - 1.053x : s_{y \cdot x}^2 = 15.50, \text{ with } 9 \text{ d.f.}\end{aligned}$$

3. For the suspect, $x = 22 - 18.73 = 3.27$, $\hat{Y} = 44.27 - (1.053)(3.27) = 40.83$, $Y = 53$.

4. Since the suspect was not used in computing this line, we can regard it as a new member of the population, and test whether its deviation from the line is within sampling error. We have $Y - \hat{Y} = 53 - 40.83 = 12.17$. Since formula 6.12.1 is applicable to the reduced sample of size $(n - 1)$, the variance due to sampling errors is

$$\begin{aligned}s_{Y - \hat{Y}}^2 &= s_{y \cdot x}^2 \left(1 + \frac{1}{n - 1} + \frac{x^2}{\Sigma x^2} \right) \\ &= (15.50) \left(1 + \frac{1}{11} + \frac{(3.27)^2}{914} \right) = (15.50)(1.1026) = 17.09\end{aligned}$$

The value of t is

$$t = \frac{Y - \hat{Y}}{s_{Y - \hat{Y}}} = \frac{12.17}{\sqrt{17.09}} = 2.943,$$

with 9 d.f. The 2% level of t is 2.821 and the 1% level is 3.250. By interpolation, P is about 0.019.

As it stands, however, this t -test does not apply, because the test assumes that the new member is randomly drawn. Instead, we selected it because it gave the largest deviation of the 12 points. If P is the probability that t for a random deviation exceeds some value t_0 , then for small values of P the probability that t_{\max} (computed for the largest of n deviations) exceeds t_0 is roughly nP . Consequently, the significance probability for our t -test is approximately $(12)(0.019) = 0.23$, and the null hypothesis is not rejected.

When the null hypothesis is rejected, this indicates an inquiry to see whether there were any circumstances peculiar to this point, or any error of measurement or recording, that caused the large deviation. In some cases an error is unearthed and corrected. In others, some extraneous causal factor that made the point aberrant is discovered, although the fault cannot be corrected. In this event, the point should be omitted in the line that is to be reported and used, provided that the causal factor is known to affect only this point. When no explanation is found the situation is perplexing. It is usually best to examine the conclusions obtained with the suspect (i) included, (ii) excluded. If these conclusions differ materially, as they sometimes do, it is well to note that either may be correct.

6.14—Prediction of X from Y . Linear calibration. In some applications the regression line is used to predict X and Y , but is constructed by measuring Y at selected values of X . In this event, as pointed out in the discussion in section 6.9 (p. 150), the prediction must be made from the regression of Y on X . For example, X may be the concentration of some element (e.g., boron or iron) in a liquid or in plant fiber and Y a quick chemical or photometric measurement that is linearly related to X . The investigator makes up a series of specimens with known amounts of X and measures Y for each specimen. From these data, the calibration curve, the linear regression of Y on X , is computed. Having measured Y for a new specimen, the estimate of $x = X - \bar{X}$ is

$$\hat{x} = (Y - \bar{Y})/b$$

Confidence limits for x and X are obtained from the method in section 6.12 by which we obtained confidence limits for Y given x . As an illustration we cite the example of sections 6.11–6.12 in which Y = percentage of wormy fruits; X = size of crop (though with these data we would in practice use the regression of X on Y , since both regressions are meaningful).

We shall find 95% confidence limits for the size of crop in a new tree with 40 per cent of wormy fruit. Turn to figure 6.11.1 (p. 155). Draw a horizontal line at $Y = 40$. The two confidence limits are the values of X at the points where this line meets the confidence curves GE and HF . Our eye readings were $X = 12$ and $X = 38$. The point estimate \hat{X} of X is, of course, the value of X , 24, at which the horizontal line meets the fitted regression line.

For a numerical solution, the fitted line is $\bar{Y} + bx$, where $\bar{Y} = 45$, $b = -1.013$. Hence the value of x when $Y = 40$ is estimated as

$$\hat{x} = (Y - \bar{Y})/b = -(40-45)/1.013 = 4.936 : \hat{X} = 23.9 \text{ hundreds}$$

To find the 95% confidence limits for x we start with the confidence limits of Y given x :

$$Y = \bar{Y} + bx \pm ts_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{x^2}{\Sigma}} \quad (6.14.1)$$

where Σ denotes Σx^2 and t is the 5% level for $(n-2)$ d.f. Expression (6.14.1) is solved as a quadratic equation in x for given Y . After some manipulation the two roots can be expressed in the following form, which appears the easiest for numerical work:

$$x = \frac{\hat{x} \pm \frac{ts_{y \cdot x}}{b} \sqrt{\frac{(n+1)}{n} (1 - c^2) + \frac{\hat{x}^2}{\Sigma}}}{1 - c^2}, \quad (6.14.2)$$

where

$$c^2 = \frac{t^2 s_b^2}{b^2} = \frac{1}{\Sigma} \left(\frac{ts_{y \cdot x}}{b} \right)^2$$

In this example $n = 12$, $t = 2.228$ (10 d.f.), $s_{y \cdot x} = 5.233$, $\Sigma = 924$, $b = -1.013$, $\hat{x} = 4.936$. These give

$$\frac{ts_{y \cdot x}}{b} = \frac{(2.228)(5.233)}{-1.013} = -11.509 ; c^2 = \frac{(11.509)^2}{924} = 0.1434$$

From (6.14.2) the limits for x are

$$x = 4.936 \pm (11.509) \sqrt{\{(1.0833)(0.8566) + 0.0264\}}$$

This gives -7.4 and $+18.9$ for x or 11.6 and 37.9 for X , in close agreement with the graphical estimate.

The quantity $c = ts_b/b$ is related to the test of significance of b . If b is significant at the 5% level, $b/s_b > t$, so that $c < 1$ and hence $c^2 < 1$. If b is not significant, the denominator in equation (6.14.2) becomes negative, and finite confidence limits cannot be found by this approach. If c is small (b highly significant), c^2 is negligible and the limits become

$$\hat{x} \pm \frac{ts_{y \cdot x}}{b} \sqrt{1 + \frac{1}{n} + \frac{\hat{x}^2}{\Sigma x^2}}$$

These are of the form $\hat{x} \pm ts_x$, where s_x denotes the factor that multiplies t . In large samples, s_x can be shown to be the estimated standard error of \hat{x} , as this result suggests.

In practice, Y is sometimes the average of m independent measurements on the new specimen. The number 1 under the square root sign in (6.14.1) then becomes $1/m$.

6.15—Partitioning the sum of squares of the dependent variate. Regression computations may be looked upon as a process of partitioning ΣY^2 into 3 parts which are both useful and meaningful. You have become accustomed to dividing ΣY^2 into $(\Sigma Y)^2/n$ and the remainder, Σy^2 ; then subdividing Σy^2 into $(\Sigma xy)^2/\Sigma x^2$ and $\Sigma d_{y \cdot x}^2$. This means that you have divided ΣY^2 into three portions:

$$\Sigma Y^2 = (\Sigma Y)^2/n + (\Sigma xy)^2/\Sigma x^2 + \Sigma d_{y \cdot x}^2$$

Each of these portions can be associated exactly with the sum of squares of a segment of the ordinates, Y . To illustrate this a simple set of data has been set up in table 6.15.1 and graphed in figure 6.15.1.

In the figure the ordinate at $X = 12$ is partitioned into 3 segments:

$$Y = \bar{Y} + \hat{y} + d_{y \cdot x},$$

where $\hat{y} = \hat{Y} - \bar{Y} = bx$ is the deviation of the point \hat{Y} on the fitted line from \bar{Y} . Each of the other ordinates may be divided similarly, though

TABLE 6.15.1
DATA SET UP TO ILLUSTRATE THE PARTITION OF ΣY^2

X	2	4	6	8	10	12	14	$\Sigma X = 56$
Y	4	2	5	9	3	11	8	$\Sigma Y = 42$
$n = 7, \bar{X} = 8, \bar{Y} = 6, \Sigma x^2 = 112, \Sigma y^2 = 68, \Sigma xy = 56$								

negative segments make the geometry less obvious. The lengths are all set out in table 6.15.2 and the several segments are emphasized in figure 6.15.1. Observe that in each line of the table (including the two at the bottom) the sum of the last three numbers is equal to the number in column Y . Corresponding to the relation

$$Y = \bar{Y} + \hat{y} + d_{y \cdot x},$$

we have the following identity in the sums of squares

$$\Sigma Y^2 = \Sigma \bar{Y}^2 + \Sigma \hat{y}^2 + \Sigma d_{y \cdot x}^2,$$

each of the three product terms being zero. The sums of squares of the ordinates, $\Sigma Y^2 = 320$, and of the deviations from regression, $\Sigma d_{y \cdot x}^2 = 40$,

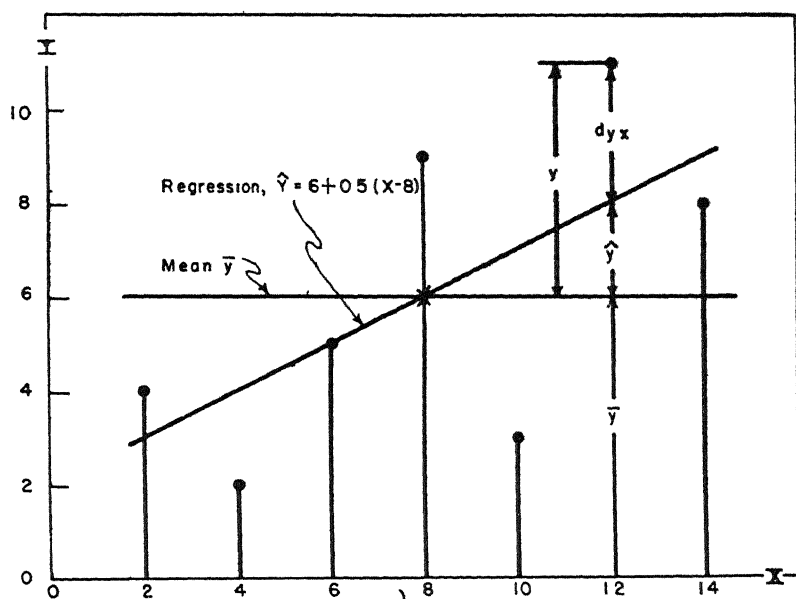


FIG 6.15.1—Graph of data in table 6.15.1. The ordinate at $X = 12$ is shown divided into 2 parts, $\bar{Y} = 6$ and $\hat{y} = 5$. Then \hat{y} is subdivided into $\hat{y} = 2$ and $d_{y \cdot x} = 3$. Thus $Y = \bar{Y} + \hat{y} + d_{y \cdot x} = 6 + 2 + 3 = 11$.

TABLE 6 15 2
LENGTHS OF ORDINATES IN TABLE 6 15 1 TOGETHER WITH
SEGMENTS INTO WHICH THEY ARE PARTITIONED

Pair Number	Ordinate Y	Mean \bar{Y}	Deviation \hat{y}	Deviation From Regression $d_{y \cdot x}$
1	4	6	-3	1
2	2	6	-2	-2
3	5	6	-1	0
4	9	6	0	3
5	3	6	1	-4
6	11	6	2	3
7	8	6	3	-1
Sum	42	42	0	0
Sum of squares	320	252	28	40

are already familiar. It remains to identify $(\Sigma Y)^2/n$ with $\Sigma \bar{Y}^2$ and $(\Sigma xy)^2/\Sigma x^2$ with $\Sigma \hat{y}^2$. First,

$$\frac{(\Sigma Y)^2}{n} = \frac{(n\bar{Y})^2}{n} = n\bar{Y}^2 = \Sigma \bar{Y}^2$$

That is, the correction for the mean is simply the sum of squares of the mean taken n times. Second,

$$\frac{(\Sigma xy)^2}{\Sigma x^2} = \frac{(\Sigma xy)^2}{(\Sigma x^2)^2} \Sigma x^2 = b^2 \Sigma x^2 = \Sigma b^2 x^2 = \Sigma \hat{y}^2$$

So the sum of squares attributable to the regression turns out to be the sum of squares of the deviations of the points \hat{Y} on the fitted line from the mean.

The vanishing of the cross-product terms is easily verified by the method used in section 6 6.

Corresponding to the partition of ΣY^2 there is a partition of the

TABLE 6 15 3
ANALYSIS OF VARIANCE OF Y IN TABLE 6 15 1

Description of Source of Variation	Symbol	Degrees of Freedom	Sum of Squares	Mean Square
The mean	\bar{Y}	1	$(\Sigma Y)^2/n = 252$	
Regression	b	1	$(\Sigma xy)^2/\Sigma x^2 = 28$	
Deviation from regression	$d_{y \cdot x}$	$n - 2 = 5$	$\Sigma d_{y \cdot x}^2 = 40$	$s_{y \cdot x}^2 = 8$
Total	Y	$n = 7$	$\Sigma Y^2 = 320$	

$$\Sigma y^2 = 28 + 40 = 68, \quad df = n - 1 = 6.$$

total degrees of freedom into three parts. Both partitions are shown in table 6 15.3. The $n = 7$ observations contribute 7 degrees of freedom, of which 1 is associated with the mean and 1 with the slope b of the regression coefficient, leaving 5 for the deviations from regression. In most applications the first line in this table is omitted as being of no interest, the breakdown taking the form presented in table 6 15.4.

TABLE 6 15 4
ANALYSIS OF VARIANCE OF y IN TABLE 6 15 1

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Regression	1	28	
Deviations from regression	5	40	8
Deviations from mean	6	68	11.3

Table 6 15 4 is an *analysis of variance* table. In addition to providing a neat summary of calculations about variability, it proves of great utility when we come to study curved regressions and comparisons among more than two means. The present section is merely an introduction to the technique, one of the major contributions of R. A. Fisher (5).

EXAMPLE 6 15 1 Dawes (6) determined the "density" of the melanin content of the skin of 24 male frogs together with their weights. Since "Some of the 24 males were selected for extreme duskiness or pallor so as to provide a measure of the extent of variability" that is, since selection was exercised on density this variate must be taken as X .

Density, X	0.13	0.15	0.28	0.58	0.68	0.31	0.35	0.58
Weight, Y	13	18	18	18	18	19	21	22
Density, X	0.03	0.69	0.38	0.54	1.00	0.73	0.77	0.82
Weight, Y	22	24	25	25	25	27	27	27
Density, X	1.29	0.70	0.38	0.54	1.08	0.86	0.40	1.67
Weight, Y	28	29	30	30	35	37	39	42

Calculate $\bar{X} = 0.6225$ units, $\bar{Y} = 25.79$ grams, $\Sigma x^2 = 3.3276$, $\Sigma y^2 = 1,211.96$, $\Sigma xy = 40.022$

EXAMPLE 6 15 2—In example 6 15 1 test the hypothesis, $\beta = 0$. Ans. $t = 3.81$, $P < 0.01$

EXAMPLE 6 15 3—Analyze the variance of the frog weights, as follows

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Mean	1	15,965.04	
Regression	1	481.36	
Deviations	22	730.60	33.21
Total	24	17,177.00	

EXAMPLE 6.15.4—How nearly free from error is the measurement of melanin density, X' ? After preparation of a solution from the skin of the frogs, the intensity of the color was evaluated in a colorimeter and the readings then transferred graphically into neutral densities. The figures reported are means of from 3 to 6 determinations. The error of this kind of measurement is usually appreciable. This makes the estimate of regression biased downwards. Had not the investigator wished to learn about extremes of density, the regression of density on weight might have been not only unbiased but more informative.

6.16—Galton's use of the term "regression." In his studies of inheritance Galton developed the idea of regression. Of the "law of universal regression" (7) he said, "Each peculiarity in a man is shared by his kinsman, but *on the average* in a less degree." His friend, Karl Pearson (8), collected more than a thousand records of heights of members of family groups. Figure 6.16.1 shows his regression of son's height on

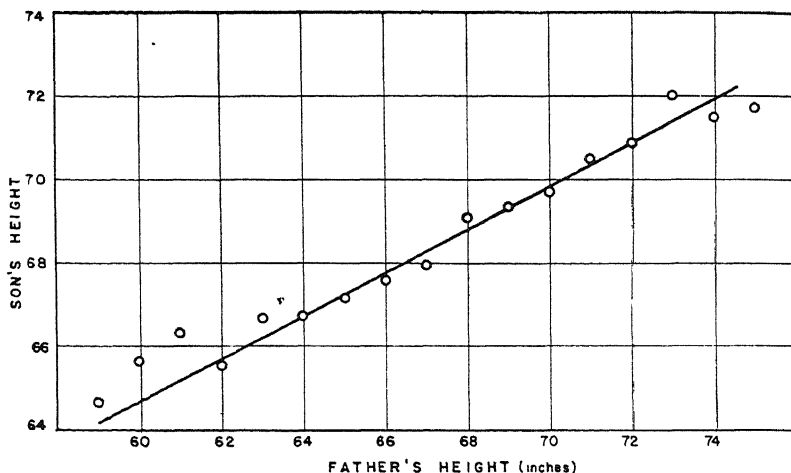


FIG 6.16.1—Regression of son's stature on father's (8) $\hat{Y} = 0.516X + 33.73$
1,078 families

father's. Though tall fathers do tend to have tall sons, yet the average height of sons of a group of tall fathers is less than their father's height. There is a *regression*, or going back, of son's heights toward the average height of all men, as evidenced by the regression coefficient, 0.516, substantially less than 1.

6.17—Regression when X is subject to error. Thus far we have assumed that the X -variable in regression is measured without error. Since no measuring instrument is perfect, this assumption is often unrealistic. A more realistic model is one that assumes $Y = \alpha + \beta(X - \bar{X}) + \varepsilon$ as before, but regards X as an *unknown* true value. Our measurement of X is $X' = X + e$, where e is the error of measurement. For any specimen we know (Y, X') but not X .

If the measurement is unbiased, e , like ε , is a random variable follow-

ing a distribution with mean 0. The errors e may arise from several sources. For instance, if X is the average price of a commodity or the average family income in a region of a country, this is usually estimated from a sample of shops or of families, so that X' is subject to a sampling error. With some concepts like "educational level" or "economic status" there may be no fully satisfactory method of measurement, so that e may represent in part measurement of the wrong concept.

If ε , e , and the true X are all normally and independently distributed it is known that Y and X' follow a bivariate normal distribution (section 7.4.). The regression of Y on X' is linear, with regression coefficient

$$\beta' = \beta/(1 + \lambda),$$

where $\lambda = \sigma_e^2/\sigma_X^2$. (If X is not normal, this result holds in large samples and approximately in small samples if λ is small.) Thus, with errors in X , the sample regression coefficient, b' , of Y on X' no longer provides an unbiased estimate of β , but of $\beta/(1 + \lambda)$.

If the principal objective is to estimate β , often called the *structural* regression coefficient, the extent of this distortion downwards is determined by the ratio $\lambda = \sigma_e^2/\sigma_X^2$. Sometimes it is possible to obtain an estimate s_e^2 of σ_e^2 . Since $\sigma_{X'}^2 = \sigma_X^2 + \sigma_e^2$, an estimate of λ is $\hat{\lambda} = s_e^2/(\sigma_{X'}^2 - s_e^2)$. From $\hat{\lambda}$ we can judge whether the downward bias is negligible or not. If it is not negligible, the revised estimate $b'(1 + \hat{\lambda})$ should remove most of the bias.

In laboratory experimentation, λ is often small even with a measuring instrument that is not highly accurate. For example, suppose that $\sigma_X = 20$, $\mu_X = 100$, so that nearly all the values of the true X 's lie between 50 and 150. Consider $\sigma_e = 3$. This implies that about half of the true X 's are measured with an error greater than 2 and about one third of them with an error greater than 3—a rather imprecise standard of performance. Nevertheless, λ is only $9/400 = 0.022$.

If the objective is to predict the population regression line or the value of an individual Y from the sample of values (Y, X') , the methods of sections 6.11 and 6.12 may still be used, with X' in place of X , provided that X , e , and ε are approximately normal. The presence of errors in X decreases the accuracy of the predictions, because the residual variance is increased, though to a minor extent if λ is small. The relation between $\sigma_{Y \cdot X'}^2$ and $\sigma_{Y \cdot X}^2$ may be put in two equivalent forms:

$$\sigma_Y^2 - \sigma_{Y \cdot X'}^2 = (\sigma_Y^2 - \sigma_{Y \cdot X}^2)/(1 + \lambda), \quad (6.17.1)$$

or,

$$\sigma_{Y \cdot X}^2 = \sigma_{Y \cdot X'}^2 + \frac{\lambda}{(1 + \lambda)} (\sigma_Y^2 - \sigma_{Y \cdot X}^2) \quad (6.17.2)$$

Berkson (10) has pointed out an exception to the above analysis. Many laboratory experiments are conducted by setting X' at a series of *fixed* values. For instance, a voltage may be set at a series of prede-

terminated levels X_1', X_2', \dots on a voltmeter. Owing to errors in the voltmeter or other defects in the apparatus, the true voltages X_1, X_2, \dots differ from the set voltages.

In this situation we still have $Y = \alpha + \beta X + \varepsilon$, $X' = X + e$. In both our original case (X normal) and in Berkson's case (X' fixed) it follows that

$$Y = \alpha + \beta X' + (\varepsilon - \beta e) \quad (6.17.3)$$

The difference is this. In our case, e and X' are correlated because of the relation $X' = X + e$. Consequently, the residual $(\varepsilon - \beta e)$ is correlated with X' and does not have a mean zero for fixed X' . This vitiates Assumption 2 of the basic model (section 6.4). With X' fixed, however, e is correlated with X but not with X' , and the model (6.17.3) satisfies the assumptions for a linear regression. The important practical conclusion is that b' , the regression of Y on X' , remains an unbiased estimate of β .

6.18—Fitting a straight line through the origin. From some data the nature of the variable Y and X makes it clear that when $X = 0$, Y must be 0. If a straight line regression appears to be a satisfactory fit, we have the relation

$$Y = \beta X + \varepsilon$$

where, in the simplest situations, the residual ε follows $\mathcal{N}(0, \sigma^2)$. The least squares estimate of β is $b = \Sigma XY / \Sigma X^2$. The residual mean square is

$$s_{y \cdot x}^2 = \{\Sigma Y^2 - (\Sigma XY)^2 / \Sigma X^2\} / (n - 1)$$

with $(n - 1)$ d.f. Confidence limits for β are

$$b \pm t s_b,$$

where t is read from the t -table with $(n - 1)$ d.f. and the appropriate probability.

This model should not be adopted without careful inspection of the data, since complications can arise. If the sample values of X are all some distance from zero, plotting may show that a straight line through the origin is a poor fit, although a straight line that is not forced to go through the origin seems adequate. The explanation may be that the population relation between Y and X is curved, the curvature being marked near zero but slight in the range within which X has been measured. A straight line of the form $(a + bx)$ will then be a good approximation within the sample range, though untrustworthy for extrapolation. If the mathematical form of the curved relation is known, it may be fitted by methods outlined in chapter 15.

It is sometimes useful to test the null hypothesis that the line, assumed straight, goes through the origin. The first step is to fit the usual two-parameter line $(\alpha + \beta x)$, i.e., $\alpha + \beta(X - \bar{X})$, by the methods given earlier in this chapter. The condition that the population line goes

through the origin is $\alpha - \beta\bar{X} = 0$. The sample estimate of this quantity is $\bar{Y} - b\bar{X}$, with estimated variance

$$s_{y \cdot x}^2 (1/n + \bar{X}^2/\Sigma x^2)$$

Hence, the value of t for the test of significance is

$$t = \frac{\bar{Y} - b\bar{X}}{s_{y \cdot x} \sqrt{\{1/n + \bar{X}^2/\Sigma x^2\}}} \quad (6.18.1)$$

with $(n - 2)$ *d.f.* This test is a particular case of the technique presented in section 6.11 for finding confidence limits for the population mean value of Y corresponding to a given value of X .

The following example comes from a study (9) of the forces necessary to draw plows at the speeds commonly attained by tractors. Those results of the regression calculations that are needed are shown under table 6.18.1.

TABLE 6.18.1
DRAFT AND SPEED OF PLOWS DRAWN BY TRACTORS

Draft (lbs.)	Y	425	420	480	495	540	530	590	610	690	680
Speed (m.p.h.)	X	0.9	1.3	2.0	2.7	3.4	3.4	4.1	5.2	5.5	6.0

$\bar{X} = 3.45$ m.p.h.	$\bar{Y} = 546$ lbs.	$n = 10$
$\Sigma x^2 = 27.985$	$\Sigma y^2 = 82,490$	$\Sigma xy = 1,492.0$
$b = 53.31$ lbs. per mile		
$s_{y \cdot x}^2 = 368.1$ with 8 <i>d.f.</i>		

One might suggest that the line should go through the origin, since when the plow is not moving there is no draft. However, inspection of table 6.18.1, or a plot of the points, makes it clear that when the line is extrapolated to $X = 0$, the predicted \bar{Y} is well above 0, as would be expected since inertia must be overcome to get the plow moving. From (6.18.1) we have

$$t = \frac{546 - (53.34)(3.45)}{\sqrt{\left[(368.1) \left\{ \frac{1}{10} + \frac{(3.45)^2}{27.985} \right\} \right]}} = \frac{362.0}{13.90} = 26.0$$

with 8 *d.f.*, confirming that the line does not go through the origin.

When the line is straight and passes through $(0, 0)$, the variance of the residual ε is sometimes not constant, but increases as X moves away from zero. On plotting, the points lie close to the line when X is small but diverge from it as X increases. The extension of the method of least squares to this case gives the estimate $b = \Sigma w_x XY / \Sigma w_x X^2$, where w_x is the reciprocal of the variance of ε at the value of X in question.

If numerous observations of Y have been made at each selected X , the variance of ε can be estimated directly for each X and the form of the

functions w_x determined empirically. If there are not enough data to use this method, simple functions that seem reasonable are employed. A common one when all X 's are positive is to assume that the variance of ε is proportional to X , so that $w_x = k/X$, where k is a constant. This gives the simple estimate $b = \Sigma Y/\Sigma X = \bar{Y}/\bar{X}$. The weighted mean square of the residuals from the fitted line is

$$s_{y \cdot x}^2 = \{\Sigma(Y^2/X) - (\Sigma Y)^2/\Sigma X\}/(n - 1)$$

and the estimated standard error of b is $s_{y \cdot x}/\sqrt{\Sigma X}$.

TABLE 6.18.2
NUMBER OF ACRES IN CORN ON 25 FARMS IN SOUTH DAKOTA (1942)
SELECTED BY FARM SIZE

Size of Farm (acres) X	Acres in Corn Y	Range	Standard Deviation s_y	Ratio s_y/X	Ratio Y/X
80	25				0.312
	10				.125
	20				.250
	32				.400
	20	22	8.05	0.101	.250
160	60				0.375
	35				.219
	20				.125
	45				.281
	40	40	14.58	0.091	.250
240	65				0.271
	80				.333
	65				.271
	85				.354
	30	55	21 51	0.090	.125
320	70				0.219
	110				.344
	30				.094
	55				.172
	60	80	29 15	0 091	.188
400	75				0.188
	35				.088
	140				.350
	90				.225
	110	105	39.21	0 098	.275
Mean	56 28				0 243

$$n = 25, b = \frac{\Sigma(Y/X)}{n} = 0.243 \text{ corn acre/farm acre}$$

Sometimes the *standard deviation* of ϵ is proportional to X , so that $w_x = k/X^2$. This leads to the least squares estimate

$$b = \Sigma(XY/X^2)/\Sigma(X^2/X^2) = \Sigma(Y/X)/n,$$

in other words, the mean of the individual ratios Y/X . This model is illustrated by the data in table 6.18.2, taken from a farm survey in eastern South Dakota in 1942, in which the size of the farm X and the number of acres in corn Y were measured. Five of the commoner farm sizes: 80, 160, 240, 320, and 400 acres, were drawn. For each size, five farm records were drawn at random.

The ranges of the several groups of Y indicate that σ is increasing with X . The same thing is shown in figure 6.18.1. To get more detailed information, s_y was calculated for each group, then the ratio of s_y to X . These ratios are so nearly constant as to justify the assumption that in the population σ_y/X is a constant. Also it seems reasonable to suppose that $O(0, 0)$ is a point on the regression line.

The value of b , 0.243 corn acres per farm acre, is computed in table 6.18.2 as the mean of the ratios Y/X . The sample regression line is $\hat{Y} = 0.243X$.

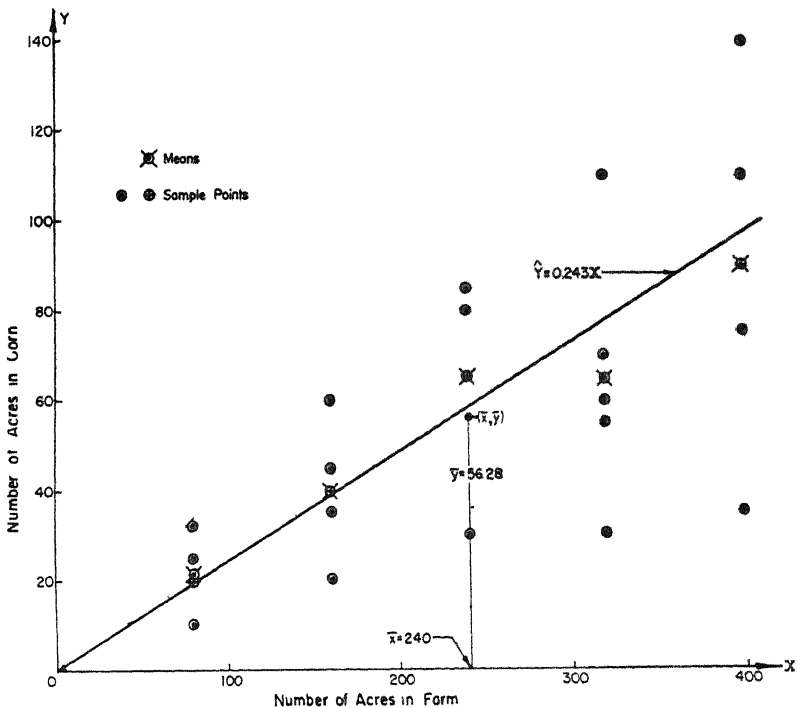


FIG 6 18 1—Regression of corn acres on farm acres

To find the estimated variance of b , first compute the sum of squares of deviations of the 25 ratios $R = Y/X$ from their means, and divide by $n - 1 = 24$. This gives $s_R^2 = 0.008069$. Then

$$s_b^2 = \frac{s_R^2}{n} = \frac{0.008069}{25} = 0.0003228$$

$$s_b = 0.0180, \text{ d.f.} = n - 1 = 24.$$

The 95% interval estimate of β is set in the usual way,

$$b - t_{0.05} s_b \leq \beta \leq b + t_{0.05} s_b,$$

the result being $0.206 \leq \beta \leq 0.280$.

In straight lines through the origin the point (\bar{X}, \bar{Y}) does not in general lie on the fitted line. In the figure, (240, 56.28) falls below the line. An exception occurs when σ_ε^2 is proportional to X , giving $b = \bar{Y}/\bar{X}$ as we have seen.

6.19—The estimation of ratios. With data in which it is believed that Y is proportional to X , apart from sampling or experimental error, the investigator is likely to regard his objective as that of estimating the common ratio Y/X rather than as a problem in regression. If his conjecture is correct, that is, if $Y = \beta X + \varepsilon$, the three quantities $\Sigma XY/\Sigma X^2$, $\Sigma Y/\Sigma X$ and $\Sigma(Y/X)/n$ are all unbiased estimates of the population ratio β . The choice among the three is a question of precision. The most precise estimate is the first, second, or third above according as the variance of ε is constant, proportional to X , or proportional to X^2 . If the variance of ε is expected to increase moderately as X increases, though the exact rate is not known, the estimate $\Sigma Y/\Sigma X$ usually does well, in addition to being the simplest of the three.

Before one of these estimates is adopted, always check that Y is proportional to X by plotting the data and, if necessary, testing the null hypothesis that the line goes through the origin. Hasty adoption of some form of ratio estimate may lose the information that Y/X is not constant as X varies.

6.20—Summary. The six sample values, \bar{n} , \bar{X} , \bar{Y} , Σx^2 , Σy^2 , Σxy , furnish all regression information about the population line $\mu = \alpha + \beta x$:

1. The regression coefficient of Y on X : $b = \Sigma xy/\Sigma x^2$. The estimate of α : $a \doteq \bar{Y}$
2. The sample regression equation of Y on X : $\hat{Y} = \bar{Y} + bx$
3. Y adjusted for X : Adjusted $Y = Y - bx$
4. The sum of squares attributable to regression:

$$(\Sigma xy)^2/\Sigma x^2 = \Sigma \hat{y}^2$$

5. The sum of squares of deviations from regression:

$$\Sigma y^2 - (\Sigma xy)^2 / \Sigma x^2 = \Sigma d_{y \cdot x}^2$$

6. The mean square deviation from regression:

$$\Sigma d_{y \cdot x}^2 / (n - 2) = s_{y \cdot x}^2$$

7. The sample standard error of \bar{Y} estimated from X :

$$s_{\bar{Y} \cdot x} = s_{y \cdot x} / \sqrt{n}$$

8. The sample standard deviation of the regression coefficient:

$$s_b = s_{y \cdot x} / \sqrt{\Sigma x^2}$$

9. The sample standard deviation of \hat{Y} as an estimate of $\mu = \alpha + \beta x$:

$$s_{\hat{Y}} = s_{y \cdot x} \sqrt{1/n + x^2 / \Sigma x^2}$$

10. The sample standard deviation of \hat{Y} as an estimate of a new point Y :

$$s_{\hat{Y}} = s_{y \cdot x} \sqrt{1 + 1/n + x^2 / \Sigma x^2}$$

11. The estimated height of the line when $X = 0$: $\bar{Y} - b\bar{X}$. This is sometimes called the *intercept* or the *elevation* of the line.

REFERENCES

1. P. P. SWANSON, *et al.* *J. Gerontology*, 10:41 (1955).
2. J. B. WENTZ and R. T. STEWART. *J. Amer. Soc. Agron.*, 16:534 (1924)
3. T. R. HANSBERRY and C. H. RICHARDSON *Iowa State Coll. J. Sci.*, 10 27 (1935).
4. G. W. SNEDECOR and W. R. BRENNEMAN. *Iowa State Coll. J. Sci.*, 19:33 (1945).
5. R. A. FISHER. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh (1925).
6. B. DAWES. *J. Exp. Biology*, 18:26 (1946).
7. F. GALTON. *Natural Inheritance*. Macmillan, London (1889)
8. K. PEARSON and A. LEE. *Biometrika*, 2 357 (1903)
9. E. V. COLLINS *Trans. Amer. Soc. Agric. Engineers*, 14:164 (1920)
10. J. BERKSON *J. Amer. Statist. Ass.*, 45 164 (1950)
11. F. S. ACTON *Analysis of Straight Line Data*. Wiley, New York (1959)

Correlation

7.1—Introduction. The *correlation coefficient* is another measure of the mutual relationship between two variables. Table 7.1.1 and figure 7.1.1 show the heights of 11 brothers and sisters, drawn from a large family study by Pearson and Lee (1). Since there is no reason to think of one height as the dependent variable and the other as the independent variable, the heights are designated X_1 and X_2 instead of Y and X . To find the sample correlation coefficient, denoted by r , compute Σx_1^2 , Σx_2^2 , and Σx_1x_2 as in the previous chapter. Then,

$$r = \Sigma x_1x_2 / \sqrt{\{(\Sigma x_1^2)(\Sigma x_2^2)\}} = 0.558,$$

as shown under table 7.1.1. Roughly speaking, r is a quantitative expression of the commonly observed similarity among children of the same parents—the tendency of the taller sisters to have the taller brothers. In the figure, the value $r = 0.558$ reflects the propensity of the dots to lie in a band extending from lower left to upper right instead of being scattered randomly over the whole field. The band is often shaped like an ellipse, with the major axis sloping upward toward the right when r is positive.

EXAMPLE 7.1.1—Calculate $r = 1$ for the following pairs:

$$X_1: 1, 2, 3, 4, 5$$

$$X_2: 3, 5, 7, 9, 11$$

TABLE 7.1.1
STATURE (INCHES) OF BROTHER AND SISTER
(Illustration taken from Pearson and Lee's sample of 1,401 families)

Family Number	1	2	3	4	5	6	7	8	9	10	11
Brother, X_1	71	68	66	67	70	71	70	73	72	65	66
Sister, X_2	69	64	65	63	65	62	65	64	66	59	62

$$n = 11, \bar{X}_1 = 69, \bar{X}_2 = 64, \Sigma x_1^2 = 74, \Sigma x_2^2 = 66, \Sigma x_1x_2 = 39$$

$$r = \Sigma x_1x_2 / \sqrt{(\Sigma x_1^2)(\Sigma x_2^2)} = 39 / \sqrt{(74)(66)} = 0.558. \text{ Pearson and Lee's } r = 0.553$$

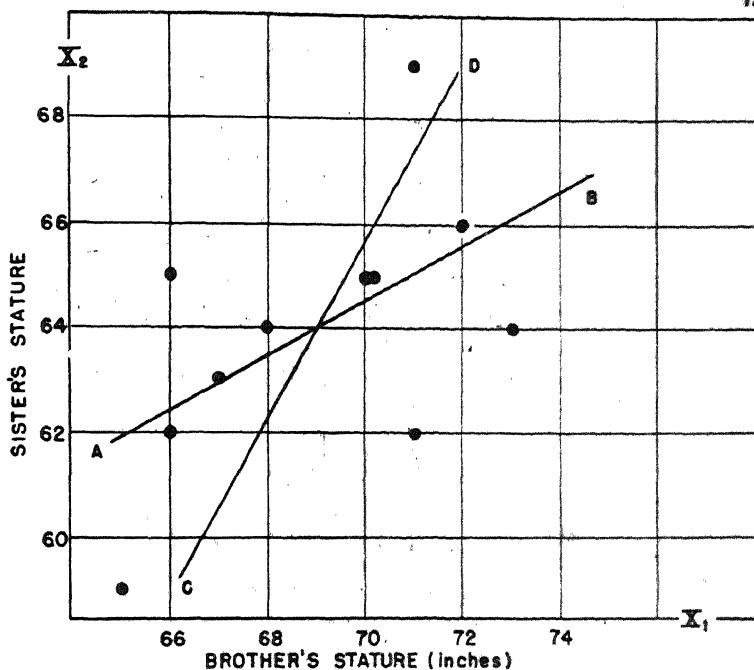


FIG. 7.1.1—Scatter (or dot) diagram of stature of 11 brother-sister pairs. $r = 0.558$.

Represent the data in a graph similar to figure 7.1.1

EXAMPLE 7.1.2—Verify $r = 0.91$ in the pairs:

X_1 : 2, 5, 6, 8, 10, 12, 14, 15, 18, 20

X_2 : 1, 2, 2, 3, 2, 4, 3, 4, 4, 5

Plot the elliptical band of points.

EXAMPLE 7.1.3—In the following, show that $r = 0.20$:

X_1 : 3, 5, 8, 11, 12, 12, 17

X_2 : 11, 5, 6, 8, 7, 18, 9

Observe the scatter of the points in a diagram.

EXAMPLE 7.1.4—In the apple data of table 6.9.1, $\Sigma x^2 = 924$, $\Sigma y^2 = 1,222$, $\Sigma xy = -936$. Calculate $r = -0.88$.

7.2—The sample correlation coefficient r . The correlation coefficient is a measure of the degree of closeness of the linear relationship between two variables.

Two properties of r should be noted:

(i) r is a pure number without units or dimensions, because the scales of its numerator and denominator are both the products of the scales in which X_1 and X_2 are measured. One useful consequence is that r can be computed from coded values of X_1 and X_2 . No decoding is required.

(ii) r always lies between -1 and $+1$ (proved in the next section, 7.3). Positive values of r indicate a tendency of X_1 and X_2 to increase together. When r is negative, large values of X_1 are associated with small values of X_2 .

To help you acquire some experience of the nature of r , a number of simple tables with the corresponding graphs are displayed in figure 7.2.1. In each of these tables $n = 9$, $\bar{X}_1 = 12$, $\bar{X}_2 = 6$, $\Sigma x_1^2 = 576$, $\Sigma x_2^2 = 144$. Only $\Sigma x_1 x_2$ changes, and with it the value of r . Since $\sqrt{(\Sigma x_1^2)(\Sigma x_2^2)} = \sqrt{(576)(144)} = 288$, the correlation is easily evaluated in the several tables by calculating $\Sigma x_1 x_2$ and dividing by 288 (or multiplying by $1/288 = 0.0034722 \dots$ if a machine is used).

In A, the nine points lie on a straight line, the condition for $r = 1$.

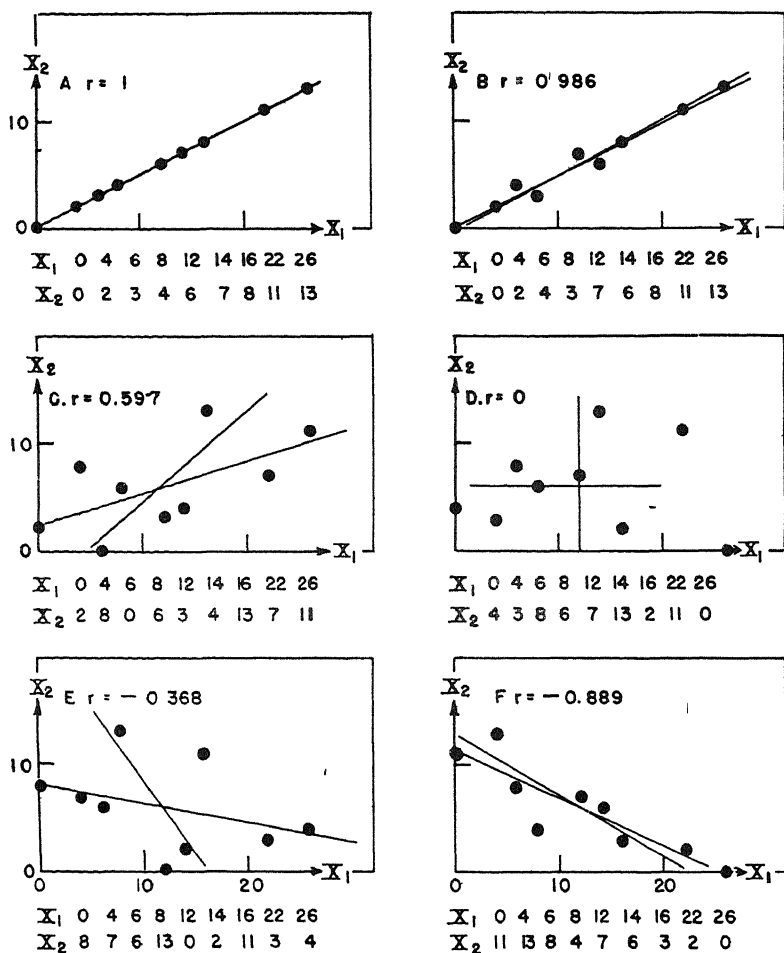


FIG 7.2.1—Scatter diagrams with correlations ranging from 1 to -0.889 .

The line is a "degenerate" ellipse—it has length but no width. The two variables keep in perfect step, any change in one being accompanied by a proportionate change in the other. *B* depicts some deviation from an exact relationship, the ellipse being long and thin with r slightly reduced below 1. In *C*, the ellipse widens, then reaches circularity in *D* where $r = 0$. This denotes no relation between the two variables. *E* and *F* show negative correlations tending toward -1 . To summarize, the thinness of the ellipse of points exhibits the magnitude of r , while the inclination of the axis upward or downward shows its sign. Note that the slope of the axis is determined by the scales of measurement adopted for the two axes of the graph and is therefore not a reliable indicator of the magnitude of r . It is the concentration of the points near the axis of the ellipse that signifies high correlation.

The larger correlations, either positive or negative, are fairly obvious from the graphs. It is not so easy to make a visual evaluation if the absolute value of r is less than 0.5; even the direction of inclination of the ellipse may elude you if r is between -0.3 and $+0.3$. In these small samples a single dot can make a lot of difference. In *D*, for example, if the point (26, 0) were changed to (26, 9), r would be increased from 0 to 0.505. This emphasizes the fact that sample correlations from a bivariate population in which the correlation is ρ are quite variable if n is small. In assessing the value of r in a table, select some extreme values of one variable and note whether they are associated with extreme values of the other. If no such tendency can be detected, r is likely small.

Perfect correlation ($r = 1$) rarely occurs in biological data, though values as high as 0.99 are not unheard of. Each field of investigation has its own range of coefficients. Inherited characteristics such as height ordinarily have correlations between 0.35 and 0.55. Among high school grades r averages around 0.35 (3). Pearson and Lee got "organic correlations," that is, correlations between two such measurements as stature and span *in the same person*, ranging from 0.60 to 0.83. Brandt (2) calculated the sample correlation, 0.986, between live weight and warm dressed weight of 533 swine. Evvard *et al.* (6) estimated $r = -0.68$ between average daily gain of swine and feed required per pound gained.

7.3—Relation between the sample coefficients of correlation and regression. If X_2 is designated as the dependent variable, its regression coefficient on X_1 , say b_{21} , is $\Sigma x_1 x_2 / \Sigma x_1^2$. But if X_1 is taken as dependent, its regression coefficient on X_2 is $b_{12} = \Sigma x_1 x_2 / \Sigma x_2^2$. The two regression lines are shown in each diagram of figure 7.2.1. The two lines are the same only if $r = \pm 1$, as illustrated in *A*, although they are close together if r is near ± 1 . In the diagrams the regression of X_1 on X_2 is always the line that makes the lesser angle with the vertical axis.

The fact that there are two different regressions is puzzling at first sight, since in mathematics the equation by which we calculate X_2 when given X_1 is the same as the equation by which X_1 is calculated when X_2

is given. In correlation and regression problems, however, we are dealing with relationships that are not followed exactly. For any fixed X_1 there is a whole population of values of X_2 . The regression of X_2 on X_1 is the line that relates the *average* of these values of X_2 to X_1 . Similarly, for each X_2 there is a population of values of X_1 , and the regression of X_1 on X_2 shows the locus of the averages of these populations as X_2 changes. The two lines answer two different questions, and coincide only if the populations shrink to their means, so that X_1 and X_2 have no individual deviation from the linear relation.

A useful property of r is obtained from the shortcut method of computing $s_{y \cdot x}^2$ in a regression problem. Reverting to Y and X , it will be recalled from the end of section 6.2 that

$$\Sigma d_{y \cdot x}^2 = (n - 2)s_{y \cdot x}^2 = \Sigma y^2 - (\Sigma xy)^2 / \Sigma x^2$$

Substituting $(\Sigma xy)^2 = r^2 \Sigma x^2 \Sigma y^2$, we have

$$\Sigma d_{y \cdot x}^2 = (n - 2)s_{y \cdot x}^2 = (1 - r^2)\Sigma y^2 \quad (7.3.1)$$

Since $\Sigma d_{y \cdot x}^2$ cannot be negative, this equation shows that r must lie between -1 and $+1$. Moreover, if r is ± 1 , $\Sigma d_{y \cdot x}^2$ is zero and the sample points lie exactly on a line.

The result (7.3.1) provides another way of appraising the closeness of the relation between two variables. The original sample variance of Y , when no regression is fitted, is $s_y^2 = \Sigma y^2 / (n - 1)$, while the variance of the deviations of Y from the linear regression is $(1 - r^2)\Sigma y^2 / (n - 2)$ as shown above. Hence, the proportion of the variance of Y that is *not* associated with its linear regression on X is estimated by

$$\frac{s_{y \cdot x}^2}{s_y^2} = \frac{(n - 1)(1 - r^2)}{(n - 2)} \doteq (1 - r^2)$$

if n is at all large. Thus r^2 may be described as *the proportion of the variance of Y that can be attributed to its linear regression on X , while $(1 - r^2)$ is the proportion free from X* . The quantities r^2 and $(1 - r^2)$ are shown in table 7 3 1 for a range of values of r .

TABLE 7 3 1
ESTIMATED PROPORTIONS OF THE VARIANCE OF Y ASSOCIATED AND
NOT ASSOCIATED WITH X IN A LINEAR REGRESSION

r	Proportion		r	Proportion	
	Associated r^2	Not ($1 - r^2$)		Associated r^2	Not ($1 - r^2$)
± 0.1	0.01	0.99	± 0.6	0.36	0.64
± 0.2	0.04	0.96	± 0.7	0.49	0.51
± 0.3	0.09	0.91	± 0.8	0.64	0.36
± 0.4	0.16	0.84	± 0.9	0.81	0.19
± 0.5	0.25	0.75	± 0.95	0.90	0.10

When r is 0.5 or less, only a minor portion of the variation in Y can be attributed to its linear regression on X . At $r = 0.7$, about half the variance of Y is associated with X , and at $r = 0.9$, about 80%. In a sample of size 200, an r of 0.2 would be significant at the 1% level, but would indicate that 96% of the variation of Y was not explainable through its relation with X . A verdict of statistical significance shows merely that there is a linear relation with non-zero slope. Remember also that convincing evidence of an association, even though close, does not prove that X is the cause of the variation in Y . Evidence of causality must come from other sources.

Another relation between the sample regression and correlation coefficients is the following. With Y as the dependent variable,

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}} \cdot \frac{\sqrt{\Sigma y^2}}{\sqrt{\Sigma x^2}} = r \frac{s_y}{s_x}$$

Or, equivalently, $r = bs_x/s_y$. Thus b is easily obtained from r , and *vice versa*, if the sample standard deviations are known.

In some applications, a common practice is to use the sample standard deviations as the scale units for measuring the variates $x = X - \bar{X}$ and $y = Y - \bar{Y}$. That is, the original variates X and Y are replaced by $x' = x/s_x$ and $y' = y/s_y$, said to be in *standard units*. The sample regression line

$$\hat{Y} - \bar{Y} = b(X - \bar{X})$$

then becomes

$$\hat{y}'s_y = bx's_x, \text{ or } \hat{y}' = \frac{bs_x}{s_y}x' = rx'$$

where \hat{y}' is the predicted value of Y in standard units. In standard measure, r is the regression coefficient, and the distinction between correlation and regression coefficients disappears.

7.4—The bivariate normal distribution. The population correlation coefficient ρ and its sample estimate r are intimately connected with a bivariate population known as the bivariate normal distribution. This distribution is illustrated by table 7.4.1 which shows the joint frequency distributions of height (X_1) and length of forearm (X_2) for 348 men. The data are from the article by Galton (18) in 1888 in which the term "co-relation" was first proposed.

To be observed in the table are five features:

(i) Each row and each column in the body of the table is a frequency distribution. Also, the column at the right, headed f_2 , is the total frequency distribution of X_2 , length of forearm, and the third-to-the-last row below is that of X_1 , height.

(ii) The frequencies are concentrated in an elliptical area with the

TABLE 7.4 1
FREQUENCY OF PAIRS OF MEASUREMENTS OF HEIGHT AND LENGTH OF FOREARM. GALTON'S DATA WITH THE OUTER CLASSES DISTRIBUTED IN PROPORTION TO TABLE 9 BY PEARSON AND LEE

Length of Forearm X_2	Height in Inches, X_1																		
	59-60	60-61	61-62	62-63	63-64	64-65	65-66	66-67	67-68	68-69	69-70	70-71	71-72	72-73	73-74	74-75	f_2	$\bar{X}_{1.2}$	$s_{1.2}$
21 0-21 5																1	1	74.4
20.5-21.0													1	1			2	72.0	0.71
20 0-20 5														1			1	72.4	..
19 5-20 0										2			1		2		5	71.0	2.51
19 0-19 5									2	4	6	11	8	4	2	1	38	70.6	1.62
18 5-19 0					1		2	6	8	7	15	13	2	1			55	68.8	1.77
18 0-18 5						3	8	15	28	14	25	5	2	2			102	68.0	1.67
17 5-18 0				2	1	2	12	18	15	7	2	1	1				61	66.7	1.62
17 0-17 5			1	3	6	11	10	7	7	3	1						49	65.4	1.80
16 5-17 0			1	5	6	5	4	1	1	1	1						25	64.4	1.97
16 0-16 5	1	1	1	3	2												8	62.0	1.41
15 5-16 0		1															1	60.4
Frequency, f_1	1	2	3	13	16	21	36	47	61	38	50	30	15	9	4	2	348		
Mean, $\bar{X}_{2.1}$	16.2	16.0	16.8	16.9	17.1	17.3	17.7	17.9	18.1	18.3	18.4	18.8	19.1	20.2	19.5	20.2	18.1		
Std Dev, $s_{2.1}$		0.36	0.50	0.52	0.60	0.48	0.54	0.42	0.51	0.68	0.48	0.41	0.70	0.83	0.29	1.41	0.905		

major axis inclined upward to the right. There are no very short men with long forearms nor any very tall men with short forearms.

(iii) The frequencies pile up along the major axis, reaching a peak near the center of the distribution. They thin out around the edges, vanishing entirely beyond the borders of the ellipse.

(iv) The center of the table is at $\bar{X}_1 = 67.5$ inches, $\bar{X}_2 = 18.1$ inches. This point happens to fall in the cell containing the greatest frequency, 28 men.

(v) The bivariate frequency histogram can be presented graphically by erecting a column over each cell in the table, the heights of columns being proportional to the cell frequencies. The tallest column would be in the center, surrounded by shorter columns. The heights would decrease toward the perimeter of the ellipse, with no columns beyond the edges. A ridge of tall columns would extend along the major axis.

The shape of the bivariate normal population becomes clear if you imagine an indefinite increase in the total frequency with a corresponding decrease in the areas of the table cells. A smooth surface would overspread the table, rising to its greatest height at the center (μ_1, μ_2), fading away to tangency with the XY plane at great distances.

Some properties of this new model are as follows:

(i) Each section perpendicular to the X_1 axis is a normal distribution, and likewise, each section perpendicular to the X_2 axis. This means that each column and each row in table 7.4.1 is a sample from a normal frequency distribution.

(ii) The frequency distributions perpendicular to the X_1 axis all have the same standard deviation, $\sigma_{2 \cdot 1}$, and they have means all lying on a straight regression line, $\mu_{2 \cdot 1} = \alpha_2 + \beta_{2 \cdot 1}X_1$. The sample means and standard deviations are recorded in the last two lines of the table. While there is considerable variation in $s_{2 \cdot 1}$, each is an estimate of the common parameter, $\sigma_{2 \cdot 1}$.

(iii) The frequency distribution perpendicular to the X_2 axis have a common standard deviation, $\sigma_{1 \cdot 2}$ (note the estimators in the right-hand column of the table), and their means lie on a second regression line, $\mu_{1 \cdot 2} = \alpha_1 + \beta_{1 \cdot 2}X_2$.

(iv) Each border frequency distribution is normal. That on the right is $\mathcal{N}(\mu_2^*, \sigma_2)$, while the one below the body of the table is $\mathcal{N}(\mu_1, \sigma_1)$.

(v) The distribution of the bivariate frequency distribution has the coefficient, $1/2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}$, followed by e with this exponent:

$$- [(X_1 - \mu_1)^2/\sigma_1^2 - 2\rho(X_1 - \mu_1)(X_2 - \mu_2)/\sigma_1\sigma_2 + (X_2 - \mu_2)^2/\sigma_2^2]/2(1 - \rho^2)$$

This distribution has five parameters. Four of them are familiar; $\mu_1, \mu_2, \sigma_1, \sigma_2$. The fifth is the correlation coefficient, ρ , of which r is an estimator. The parameter, ρ , measures the closeness of the population relation between X_1 and X_2 ; it determines the narrowness of the ellipse containing the major portion of the observations.

EXAMPLE 7.4.1—Make a graph of \bar{X}_{21} in the next-to-the-last line of table 7.4.1. The values of X_1 are the class-marks at the top of the columns. The first class mark may be taken as 59.5 inches.

EXAMPLE 7.4.2—Graph the \bar{X}_{12} on the same sheet with that of \bar{X}_{21} . The class marks for X_2 are laid off on the vertical axis. The first class mark may be taken as 21.25 inches. If you are surprised that the two regression lines are different, remember that \bar{X}_{21} is the mean of a column while \bar{X}_{12} is the mean of a row.

EXAMPLE 7.4.3—Graph s_{21} against X_1 . You will see that there is no trend, indicating that all the s_{21} may be random samples from a common σ_{21} .

EXAMPLE 7.4.4—The data in example 6.9.3 may be taken as a random sample from a bivariate normal population. You had $\bar{X} = 83$ gms., $\bar{Y} = 60$ mg., $\Sigma x^2 = 1,000$, $\Sigma y^2 = 6,854$, $\Sigma xy = 2,302$. Calculate the regression of body weight, X , on comb weight, Y . Ans. $\bar{X} = 83 + 0.336(Y - 60)$ gms. Draw the graph of this line along with that of example 6.9.4. Notice that the angle whose tangent is 0.336 is measured from the Y axis.

EXAMPLE 7.4.5—In the chick experiment, estimate $\sigma_{y \cdot x}$. Ans. $s_{y \cdot x} = 13.9$ mg. Also estimate $\sigma_{x \cdot y}$. Ans. $s_{x \cdot y} = 15.1$ gms. In $s_{x \cdot y}$, the deviations from regression are measured horizontally.

EXAMPLE 7.4.6—From the chick data, estimate ρ . Ans. $r = 0.88$.

EXAMPLE 7.4.7—If $y = a + bu$ and $x = c + dv$, where a , b , c , and d are constants, prove that $r_{xy} = r_{uv}$.

EXAMPLE 7.4.8—Thirty students scored as follows in two mathematics achievement tests:

I	73	41	83	71	39	60	51	41	85	88	44	71	52	74	50
II	29	24	34	27	24	26	35	18	33	39	27	35	25	29	13
I	43	85	53	85	44	66	60	33	43	76	51	57	35	40	76
II	13	40	23	40	22	25	21	26	19	29	25	19	17	17	35

Calculate $r = 0.774$.

From the formula for r we can derive a much used expression for ρ . Write

$$r = \Sigma(x_1 x_2) / \sqrt{\{\Sigma(x_1^2) \Sigma(x_2^2)\}}$$

Dividing both sides by $(n - 1)$, we have

$$r = \{\Sigma(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)/(n - 1)\} / s_1 s_2 \quad (7.4.1)$$

As n becomes large, \bar{X}_1 and \bar{X}_2 tend to coincide with μ_1 and μ_2 , respectively, s_1 and s_2 tend to equal σ_1 and σ_2 , and division by $(n - 1)$ becomes equivalent to division by n . Hence, when applied to the whole population, equation 7.4.1 becomes

$$\rho = \{\text{Average value of } (X_1 - \mu_1)(X_2 - \mu_2)\} / \sigma_1 \sigma_2 \quad (7.4.2)$$

The numerator of (7.4.2) is called the *population covariance* of X_1 and X_2 . This gives

$$\rho = \text{Cov.}(X_1 X_2) / \sigma_1 \sigma_2 \quad (7.4.3)$$

7.5—Sampling variation of the correlation coefficient. Common elements. A convenient way to draw samples from a normal bivariate population is by use of an old device called *common elements* (17). You may go back to the random sampling scheme of section 3.3 (p. 69), or to samples already drawn from table 3.2.1. In a new table, such as 7.5.1, record some convenient number, say three, of the random pig gains. These gains, or *elements*, are written twice in the table. Then continue the drawing, adding for example, *one* more randomly drawn gain to the left-hand column, and *two* more to the right. The sums constitute the paired values of X_1 and X_2 . Three such pairs are computed in the table. It is clear that there is a relation between the two sums in each pair. If the three common elements all happen to be large, then both X_1 and X_2 are likely large irrespective of the extra elements contained in each. Naturally, owing to the non-common elements, the relation is not perfect. If you continue

TABLE 7.5.1
CALCULATION OF THREE PAIRS OF VALUES OF THE VARIABLES X_1 AND X_2 HAVING
COMMON ELEMENTS
(The elements are pig gains from table 3.2.1)

Pair	Elements
1	$\begin{array}{ccc} 23 & & 23 \\ 44 & \leftarrow \text{common} \rightarrow & 44 \\ 43 & & 43 \\ 37 & & 30 \\ & \leftarrow \text{different} \rightarrow & 33 \\ \hline X_1 = 147 & & 173 = X_2 \end{array}$
2	$\begin{array}{ccc} 40 & & 40 \\ 16 & \leftarrow \text{common} \rightarrow & 16 \\ 19 & & 19 \\ 30 & & 29 \\ & \leftarrow \text{different} \rightarrow & 13 \\ \hline X_1 = 105 & & 117 = X_2 \end{array}$
3	$\begin{array}{ccc} 23 & & 23 \\ 38 & \leftarrow \text{common} \rightarrow & 38 \\ 37 & & 37 \\ 30 & & 31 \\ & \leftarrow \text{different} \rightarrow & 41 \\ \hline X_1 = 128 & & 170 = X_2 \end{array}$

his process, drawing a hundred or more pairs, and then compute the correlation, you will get a value of r not greatly different from the population value,

$$\rho = 3/\sqrt{(4)(5)} = 0.67$$

The numerator of this fraction is the number of common elements, while the denominator is the geometric mean of the total numbers of elements in the two sums, X_1 and X_2 . Thus, if n_{12} represents the number of common elements, with n_{11} and n_{22} designating the total numbers of elements making up the two sums, then the correlation between these two sums is, theoretically,

$$\rho = n_{12}/\sqrt{n_{11}n_{22}}$$

Of course, there will be sampling variation in the values calculated from drawings. You may be lucky enough to get a good verification with only 10 or 20 pairs of sums. With 50 pairs we have usually got a coefficient within a few hundredths of the expected parameter, but once we got 0.28 when the population was

$$n_{12}/\sqrt{n_1 n_2} = 6/\sqrt{(9)(16)} = 0.5$$

If you put the same number of elements into X_1 and X_2 , then $n_1 = n_2$. Denoting this common number of total elements by n ,

$$\rho = n_{12}/n,$$

the ratio of the number of common elements to the total number in each sum. In this special case, the correlation coefficient is simply the fraction of the elements which are common. Roughly, this is the interpretation of the sister-brother correlation in stature (table 7.1.1), usually not far from 0.5: an average of some 50% of the genes determining height is common to sister and brother.

Another illustration of this special case arises from the determination of some physical or chemical constant by two alternative methods. Consider the estimation of the potassium content of the expressed sap of corn stems as measured by two methods, the colorimetric and the gravimetric. Two samples are taken from the same source, one being treated by each of the two techniques. The common element in the two results is the actual potassium content. Extraneous elements are differences that may exist between the potassium contents of the two samples that were drawn, and the errors of measurement of the two procedures.

The concept of common elements has been presented because it may help you to a better understanding of correlation. But it is not intended as a method of interpreting the majority of the correlations that you will come across in your work, since it applies only in the type of special circumstances that we have illustrated.

When you have carried through some calculations of r with common elements, you are well aware of the sampling variation of this statistic.

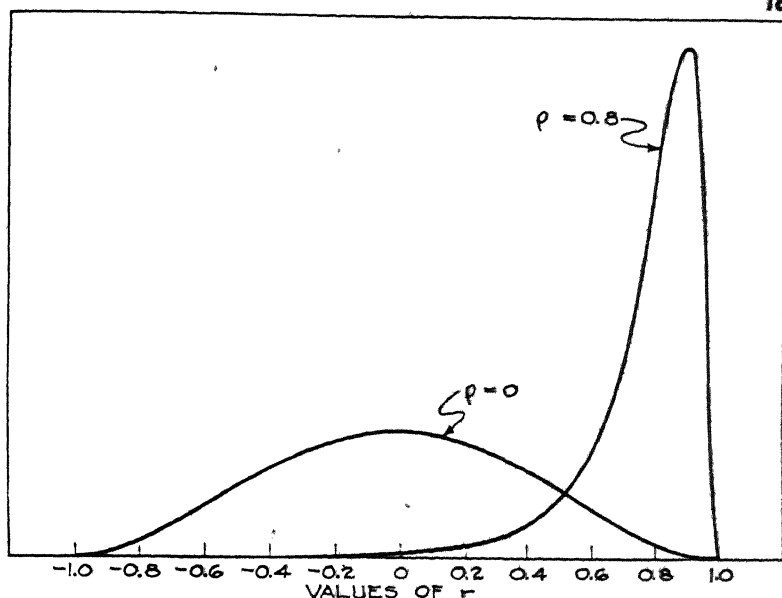


FIG. 7.5.1—Distribution of sample correlation coefficients in samples of 8 pairs drawn from two normally distributed bivariate populations having the indicated values of ρ .

However, it would be too tedious to compute enough coefficients to gain a picture of the distribution curve. This has been done mathematically from theoretical considerations. In figure 7.5.1 are the curves for samples of 8 drawn from populations with correlations zero and 0.8. Even the former is not quite normal. The reason for the pronounced skewness of the latter is not hard to see. Since the parameter is 0.8, sample values can exceed this by no more than 0.2, but may be less than the parameter value by as much as 1.8. Whenever there is a limit to the variation of a statistic at one end of the scale, with practically none at the other, the distribution curve is likely to be asymmetrical. Of course, with increasing sample size this skewness tends to disappear. Samples of 400 pairs, drawn from a population with a correlation even as great as 0.8, have little tendency to range more than 0.05 on either side of the parameter. Consequently, the upper limit, unity, would not constitute a restriction, and the distribution would be almost normal.

EXAMPLE 7.5.1—In a tea plantation (5), the production of 16 plots during one 14-week period was correlated with the production of the same plots in the following period of equal length. The correlation coefficient was 0.91. Can you interpret this in terms of common elements?

EXAMPLE 7.5.2—To prove the result that with common elements, $\rho = n_{12} \sqrt{n_{11}n_{22}}$, start from the result (7.4.3), which gives $\rho = \text{Cov.}(X_1X_2)/\sigma_1\sigma_2$. If X_1 is the sum of n_{11} independent drawings from a population with standard deviation σ , then $\sigma_1 = \sigma\sqrt{n_{11}}$. Similarly, $\sigma_2 = \sigma\sqrt{n_{22}}$. To find $\text{Cov.}(X_1X_2)$ write $X_1 = c + u_1$, $X_2 = c + u_2$, where c , the common part, is the sum of the same set of n_{12} drawings. Assuming that the drawings are from a

population with zero mean, X_1 and X_2 will have zero means. Thus, $\text{Cov.}(X_1X_2) = \text{Average value of } (X_1X_2) = \text{Average value of } (c + u_1)(c + u_2)$. But this is simply the average of c^2 , or in other words the variance of c , since the terms cu_2 , cu_1 and u_1u_2 all have averages zero because c , u_1 and u_2 result from independent drawings. Finally, the variance of c is $\sigma^2_{n_{12}}$, giving $\rho = \sigma^2_{n_{12}}/(\sigma\sqrt{n_{11}})(\sigma\sqrt{n_{22}}) = n_{12}/\sqrt{n_{11}n_{22}}$.

EXAMPLE 7.5.3—Suppose that u_1 , u_2 , u_3 are independent draws from the same population, and that $X_1 = 3u_1 + u_2$, $X_2 = 3u_1 + u_3$. What is the correlation ρ between X_1 and X_2 ? Ans 0.9. More generally, if $X_1 = fu_1 + u_2$, $X_2 = fu_1 + u_3$, then $\rho = f^2/(f^2 + 1)$. This result provides another method of producing pairs of correlated variates.

7.6—Testing the null hypothesis $\rho = 0$. From the distribution of r when $\rho = 0$, table A 11 gives the 5% and 1% significance levels of r . Note that the table is entered by the *degrees of freedom*, in this case $n - 2$. (This device was adopted because it enables the same table to be used in more complex problems.) As an illustration, consider the value $r = 0.597$ which was obtained from a sample of size 9 in diagram C of figure 7.2.1. For 7 d.f., the 5% value of r in table A 11 is 0.666. The observed r is not statistically significant, and the null hypothesis is not rejected. This example throws light on the difficulty of graphical evaluation of correlations, especially when the number of degrees of freedom is small—they may be no more than accidents of sampling. Since the distribution of r is symmetrical when $\rho = 0$, the sign of r is ignored when making the test.

Among the following correlations, observe how conclusions are affected by both sample size and the size of r :

Number of Pairs	Degrees of Freedom	r	Conclusion About Hypothesis, $\rho = 0$
20	18	0.60	Reject at 1% level
100	98	0.21	Reject at 5% level
10	8	0.60	Not rejected
15	13	-0.50	Not rejected
500	498	-0.15	Reject at 1% level

You now know two methods for testing whether there is a linear relation between the variables Y and X . The first is to test the regression coefficient $b_{y \cdot x}$ by calculating $t = b_{y \cdot x}/s_b$ and reading the t -table with $(n - 2)$ d.f. The second is the test of r . Fisher (8) showed that the two tests are identical. In fact, the table for r can be computed from the t -table by means of the relation

$$t = b_{y \cdot x}/s_b = r\sqrt{(n - 2)/\sqrt{1 - r^2}}, \quad \text{d.f.} = n - 2 \quad (7.6.1)$$

(See example 7.6.1). To illustrate, we found that the 5% level of r for 7 d.f. was 0.666. Let us compute

$$t = (0.666)\sqrt{7/\sqrt{1 - (0.666)^2}} = 2.365$$

Reference to the t -table (p. 549) shows that this is the 5% level of t for 7 d.f. In practice, use whichever test you prefer.

This relation raises a subtle point. The t -test of b requires only that Y be normally distributed: the values of X may be normal or they may be selected by the investigator. On the other hand, we have stressed that r and ρ are intimately connected with random samples from the bivariate normal distribution. Fisher proved, however, that in the particular case $\rho = 0$, the distribution of r is the same whether X is normal or not, provided that Y is normal.

EXAMPLE 7.6.1—To prove relation (7.6.1) which connects the t -test of b with the test of r , you need three relations: (i) $b_{y \cdot x} = rs_y/s_x$, (ii) $s_b = s_{y \cdot x}/\sqrt{\sum x^2}$, (iii) $s_{y \cdot x}^2 = (1 - r^2)\sum y^2/(n - 2)$, as shown in equation (7.3.1), p. 176. Start with $t = b_{y \cdot x}/s_b$ and make these substitutions to establish the result.

7.7—Confidence limits and tests of hypotheses about ρ . The methods given in this section, which apply when ρ is not zero, require the assumption that the (X, Y) or (X_1, X_2) pairs are a random sample from a bivariate normal distribution.

Table A 11 or the t -table can be used only for testing the null hypothesis $\rho = 0$. They are unsuited for testing other null hypotheses, such as $\rho = 0.5$ for example, or $\rho_1 = \rho_2$, or for making confidence statements about ρ . When $\rho \neq 0$ the shape of the distribution of r changes, becoming skew, as was seen in figure 7.5.1.

A solution of these problems was provided by Fisher (9) who devised a transformation from r to a quantity z , distributed almost normally with standard error

$$\sigma_z = \frac{1}{\sqrt{(n-3)}},$$

“practically independent of the value of the correlation in the population from which the sample is drawn.” The relation of z to r is given by

$$z = \frac{1}{2}[\log_e(1+r) - \log_e(1-r)]$$

Table A 12 (r to z) and A 13 (z to r) enable us to change from one to the other with sufficient accuracy. Following are some examples of the use of z .

1. *It is required to set confidence limits to the value of ρ in the population from which a sample r has been drawn.* As an example, consider $r = -0.889$, based on 9 pairs of observations, figure 7.2.1F. From table A 12, $z = 1.417$ corresponds to $r = 0.889$. Since $n = 9$, $\sigma_z = 1/\sqrt{6} = 0.408$. Since z is distributed almost normally, independent of sample size, $z_{0.01} = 2.576$. For $P = 0.99$, we have as confidence limits for z ,

$$1.417 - (2.576)(0.408) \leq z \leq 1.417 + (2.576)(0.408), \\ 0.366 \leq z \leq 2.468$$

Using table A 13 to find the corresponding r , and restoring the sign, the 0.99 confidence limits for ρ are given by

$$-0.986 \leq \rho \leq -0.350$$

Emphasis falls on two facts: (i) in small samples the estimate, r , is not very reliable; and (ii) the limits are not equally spaced on either side of r , a consequence of its skewed distribution.

2. *Occasionally, there is reason to test the hypothesis that ρ has some particular value, other than zero, in the sampled population* ($\rho = 0$, you recall, is tested by use of table A 11). An example was given in section 7.5, where $r = 0.28$ was observed in a sample of 50 pairs from $\rho = 0.5$. What is the probability of a larger deviation? For $r = 0.28$, $z = 0.288$, and for $\rho = 0.5$, $z = 0.549$. The difference, $0.549 - 0.288 = 0.261$, has a standard error, $1/\sqrt{(n-3)} = 1/\sqrt{47} = 0.1459$. Hence, the normal deviate is $0.261/0.1459 = 1.80$, which does not reach the 5% level: the sample is not as unusual as a 1-in-20 chance.

3. *To test the hypothesis that two sample values of r are drawn at random from the same population*, convert each to z , then test the significance of the difference between the two z 's. For two lots of pigs the correlations between gain in weight amount of feed eaten are recorded in table 7.7.1. The difference between the z -values, 0.700, has the mean square

$$\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} = \frac{1}{2} + \frac{1}{9} = 0.611$$

The test is completed in the usual manner, calculating the ratio of the difference of the z 's to the standard error of this difference. With $P = 0.37$ there is no reason to reject the hypothesis that the z 's are from the same population, and hence that the r 's are from a common population correlation.

4. *To test the hypothesis that several r 's are from the same ρ , and to combine them into an estimate of ρ* . Several sample correlations may possibly be drawn from a common ρ . If this null hypothesis is not rejected, we may wish to combine the r 's into an estimate of ρ more reliable than that afforded by any of the separate r 's. Lush (14) was interested in an average of the correlations between initial weight and gain in 6 lots of steers. The computations are shown in table 7.7.2. Each z is weighted (multiplied) by the reciprocal of its mean square, so that small samples

TABLE 7.7.1
TEST OF SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO CORRELATIONS OF GAIN
WITH FEED EATEN AMONG SWINE

Lot	Pigs in Lot	r	z	$1/(n-3)$
1	5	0.870	1.333	0.500
2	12	0.560	0.633	0.111
Difference = 0.700				Sum = 0.611
$\sigma_{z_1 - z_2} = \sqrt{0.611} = 0.782$ $0.700/0.782 = 0.895$ $P = 0.37$				

TABLE 7.7 2
TEST OF HYPOTHESIS OF COMMON ρ AND ESTIMATION OF ρ . CORRELATION BETWEEN
INITIAL WEIGHT AND GAIN OF STEERS

Samples	No. = n	$n - 3$	r	z	Weighted z = $(n - 3)z$	Weighted Square = $(n - 3)z^2$	Cor- rected z
1927 Herefords	4	1	0.929	1.651	1.651	2.726	1.589
1927 Brahmans	13	10	0.570	0.648	6.480	4.199	0.633
1927 Backcrosses	9	6	0.455	0.491	2.946	1.446	0.468
1928 Herefords	6	3	-0.092	-0.092	-0.276	0.025	-0.055
1928 Brahmans	11	8	0.123	0.124	0.992	0.123	0.106
1928 Backcrosses	14	11	0.323	0.335	3.685	1.234	0.321
57		39			15.478	9.753	14.941
				Average $\bar{z}_w = 0.397$		6.145	$z = 0.383$
Average $r = 0.377$				$\chi^2 = 3.608$		$r = 0.365$	

have little weight. The sum of the weighted z 's, 15.478, is divided by the sum of the weights, 39, to get the average $\bar{z}_w = 0.397$.

The next column contains the calculations that lead to the test of the hypothesis that the six sample correlations are drawn from a common population correlation. The test is based on a general result that if the k normal variates z_i are all estimates of the same mean μ , but have different variances σ_i^2 , then

$$\sum w_i (z_i - \bar{z}_w)^2 = \sum w_i z_i^2 - (\sum w_i z_i)^2 / \sum w_i$$

is distributed as χ^2 with $(k - 1)$ d.f., where $w_i = 1/\sigma_i^2$. In this application, $w_i = n_i - 3$ and

$$\begin{aligned}\chi^2 &= \sum (n - 3)z^2 - [\sum (n - 3)z]^2 / \sum (n - 3) \\ &= 9.753 - (15.478)^2 / 39 = 3.610,\end{aligned}$$

with 5 degrees of freedom. From table A 5, p. 550, $P = 0.61$, so that H_0 is not rejected.

Since the six sample correlations may all have been drawn from the same population, we compute an estimate of the common ρ . This is got by reading from table A 13 the correlation 0.377 corresponding to the average $\bar{z}_w = 0.397$. Don't fail to note the great variation in these small sample correlations. The S.D. of \bar{z}_w is $1/\sqrt{39}$.

Fisher pointed out that there is a small bias in z , each being too large by

$$\frac{\rho}{2(n - 1)}$$

The bias may usually be neglected. It might be serious if large numbers of correlations were averaged, because the bias accumulates, one bit being

added with every z . If there is need to increase accuracy in the calculation of table 7.7.2, the average $r = 0.377$ may be substituted for ρ ; then the approximate bias for each z may be deducted, and the calculation of the average z repeated. Since this will decrease the estimated r , it is well to guess ρ slightly less than the average r . For instance, it may be guessed that $\rho = 0.37$, then the correction in the first z is $0.37/2(4 - 1) = 0.062$, and corrected z is $1.651 - 0.062 = 1.589$. The other corrected z 's are in the last column of the table. The sum of the products,

$$\Sigma(n - 3)(\text{corrected } z) = 14.941,$$

is divided by 39 to get the corrected mean value of z , 0.383. The corresponding correlation is 0.365.

For tables of the distribution of r when $\rho \neq 0$, see reference (4).

EXAMPLE 7.7.1—To get an idea of how the selection of pairs affects correlation, try picking the five lowest values of test II (example 7.4.8) together with the six highest. The correlation between these 11 scores and the corresponding scores on test I turns out to be 0.89, as against $r = 0.77$ for the original sample.

EXAMPLE 7.7.2—Set 95% confidence limits to the correlation, 0.986, $n = 533$, between live and dressed weights of swine. Ans. 0.983 – 0.988.

What would have been the confidence limits if the number of swine had been 25? Ans. 0.968 – 0.994.

EXAMPLE 7.7.3—In four studies of the correlation between wing and tongue length in bees, Grout (10) found values of $r = 0.731, 0.354, 0.690$, and 0.740 , each based on a sample of 44. Test the hypothesis that these are samples from a common ρ . Ans. $\chi^2 = 9.164$, $d.f. = 3$, $P = 0.03$. In only about three trials per 100 would you expect such disagreement among four correlations drawn from a common population. One would like to know more about the discordant correlation, 0.354, before drawing conclusions.

EXAMPLE 7.7.4—Estimate ρ in the population from which the three bee correlations, 0.731, 0.690, and 0.740, were drawn. Ans. 0.721.

EXAMPLE 7.7.5—Set 99% confidence limits on the foregoing bee correlation. Note: $r = 0.721$ is based on $(n - 3) = 3 \times 41 = 123$. The value of z is therefore equivalent to a single z from a sample of $123 + 3 = 126$ bees. The confidence limits are: 0.590 – 0.815.

7.8—Practical utility of correlation and regression. Over the last forty years, investigators have tended to increase their use of regression techniques and decrease their use of correlation techniques. Several reasons can be suggested. The correlation coefficient r merely estimates the degree of closeness of linear relationship between Y and X , and the meaning of this concept is not easy to grasp. To ask whether the relation between Y and X is close or loose may be sufficient in an early stage of research. But more often the interesting questions are: How much does Y change for a given change in X ? What is the shape of the curve connecting Y and X ? How accurately can Y be predicted from X ? These questions are handled by regression techniques.

Secondly, the standard results for the distribution of r as an estimate of a non-zero ρ require random sampling from a bivariate normal population. Selection of the values of X at which Y is measured, often done in-

tentionally or because of operational restrictions, can distort the frequency distribution of r to a marked degree.

The correlation between two variables may be due to their common relation to other variables. The organic correlations already mentioned are examples. A big animal tends to be big all over, so that two parts are correlated because of their participation in the general size. Over a period of years, many apparently unrelated variables rise or fall together within the same country or even in different countries. There is a correlation of -0.98 between the annual birthrate in Great Britain, from 1875 to 1920, and the annual production of pig iron in the United States. The matter was discussed by Yule (19) as a question: Why do we sometimes get nonsense-correlations between time series? Social, economic, and technological changes produce the time trends that lead to such examples.

In some problems the correlation coefficient enters naturally and usefully. Correlation has played an important part in biometrical genetics, because many of the consequences of Mendelian inheritance, and later developments from it, are expressed conveniently in terms of the correlation between related persons or animals.

A second example occurs when we are trying to select persons with high values of some skill Y by means of examination results X . If Y and X follow the bivariate normal distribution, the average Y value, say Y' , of candidates whose exam score is X is given by the equation

$$(Y' - \mu_Y)/\sigma_Y = \rho(X - \mu_X)/\sigma_X$$

Suppose we select the top $P\%$ in the exam. For the normal curve, the average value of $(X - \mu_X)/\sigma_X$ for the selected men may be shown to be H/P when there are many candidates, where H is the ordinate of the normal curve at the point that separates the top $P\%$ from the remaining $(1 - P)\%$. When $P = 5\%$, the ordinate $H = 0.1032$, and $H/P = 2.06$. Thus the average Y' value of the top 5% is 2.06ρ in standard units. If $\rho = 0.5$ this average is 1.03 . From the normal tables we find that when $H/P = 1.03$, the corresponding P is 36% . This means that with $\rho = 0.5$, the 5% most successful performers in the exam have only the same average ability as the top 36% of the original candidates. The size of ρ is the key factor in determining how well we can select high values of Y by a screening process based on X .

In hydrology, suppose that there are annual records Y of the flow of a stream for a relatively short period of m years, and records X of a neighboring stream for a longer period of n years. Instead of using \bar{Y}_m as the estimate of the long-term mean μ_Y of Y , we might work out the regression of Y on X and predict μ_Y by the formula

$$\hat{\mu}_Y = \bar{Y}_m + b(\bar{X}_n - \bar{X}_m)$$

The proportional reduction in variance due to this technique, known as *stream extension*, is approximately

$$\frac{V(\bar{Y}_m) - V(\hat{\mu}_Y)}{V(\bar{Y}_m)} \doteq \frac{(n-m)}{n} \left[\rho^2 - \frac{(1-\rho^2)}{m-3} \right]$$

Here again it is the value of ρ , along with the lengths of run available in the two streams, that determines whether this technique gives worthwhile gains in precision.

7.9—Variances of sums and differences of correlated variables. When X_1 and X_2 are independent, a result used previously is that the variance of their sum is the sum of their variances. When they are correlated, the more general result is

$$\sigma_{X_1+X_2}^2 = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2 \quad (7.9.1)$$

Positive correlation increases the variance of a sum, negative correlation decreases it. The corresponding sample result is

$$s_{X_1+X_2}^2 = s_1^2 + s_2^2 + 2rs_1s_2 \quad (7.9.2)$$

This identity is occasionally used as a check on the computation of s_1 , s_2 , and r from a sample. For each member of the sample, $X_1 + X_2$ is written down and the sample variance of this quantity is obtained in the usual way.

For the difference $D = X_1 - X_2$, the variance is

$$\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 \quad (7.9.3)$$

With differences, positive correlations decrease the variance. In paired experiments, the goal in pairing is to produce a positive correlation ρ between the members X_1 , X_2 of a pair. The pairing does not affect the term $(\sigma_1^2 + \sigma_2^2)$ in (7.9.3), but brings in a negative term, $2\rho\sigma_1\sigma_2$.

If we have k variates, with ρ_{ij} the correlation between the i th and the j th variates, their sum $S = X_1 + X_2 + \dots + X_k$ has variance

$$\begin{aligned} \sigma_S^2 = & \sigma_1^2 + \sigma_2^2 + \dots + \sigma_k^2 + 2\rho_{12}\sigma_1\sigma_2 + 2\rho_{13}\sigma_1\sigma_3 + \dots \\ & + 2\rho_{k-1,k}\sigma_{k-1}\sigma_k \end{aligned} \quad (7.9.4)$$

where the cross-product terms $2\rho_{ij}\sigma_i\sigma_j$ extend over every pair of variates.

EXAMPLE 7.9.1—To prove formula (7.9.1), note that by the definition of a variance, the variance of $X_1 + X_2$ is the average value of $(X_1 + X_2 - \mu_1 - \mu_2)^2$, taken over the population. Write this as

$$E\{(X_1 - \mu_1) + (X_2 - \mu_2)\}^2 = E(X_1 - \mu_1)^2 + E(X_2 - \mu_2)^2 + 2E(X_1 - \mu_1)(X_2 - \mu_2)$$

where the symbol E (expected value) stands for "the average value of" This gives

$$\sigma_{X_1+X_2}^2 = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2,$$

since by equation (7.4.2) (p. 180), $E(X_1 - \mu_1)(X_2 - \mu_2) = \rho\sigma_1\sigma_2$. Formulas (7.9.3) and (7.9.4) are proved in the same way.

EXAMPLE 7.9.2—In a sample of 300 ears of corn (7), the weight of the grain, G , had a standard deviation $s_g = 24.62$ gms.; the weight of the cob, C , had a standard deviation $s_c = 4.19$ gms.; and r_{gc} was 0.6906. Show that the total ear weight $W = G + C$ had $s_w = 27.7$ gms. and that $r_{wg} = 0.994$.

EXAMPLE 7.9.3—In table 7.1.1, subtract each sister's height from her brother's, then compute the corrected sum of squares of the differences. Verify by formula (7.9.3) that your result agrees with the values $\Sigma x_1^2 = 74$, $\Sigma x_2^2 = 66$, $\Sigma x_1 x_2 = 39$, given under table 7.1.1

EXAMPLE 7.9.4—If $r_{12} = 1$, show that $s_D = s_1 - s_2$, where $s_1 \geq s_2$.

7.10—The calculation of r in a large sample. When the sample is large, the variates X and Y are often grouped into classes, as illustrated in table 7.10.1 for a sample of 327 ears of corn (20). The diameters X are in millimeter classes and the weights Y in 10-gram classes. The figures in the body of the table are the frequencies f_{xy} in each X and Y class. Looking at the class with diameter 48 and weight 300, we see that there were $f_{xy} = 3$ ears in this class, i.e., with diameters between 47.5 and 48.5 mm., and weights between 295 and 305 gms. Correlation in these data is evidenced by the tendency of high frequencies to lie along the diagonal of the table, leaving two corners blank—there are no very heavy ears with small diameters.

The steps in the calculation are as follows:

1. Add the frequencies in each row, giving the column of values f_y , and in each column, giving the row of values f_x .
2. Construct a convenient coding of the weights and diameters, writing down the coded Y and X values.
3. Write down a column of the values Yf_y and a row of the values Xf_x .
4. The quantities ΣXf_x , ΣYf_y , Σx^2 and Σy^2 are now found on the calculating machine in the usual way, and are entered in table 7.10.2
5. The device for finding Σxy is new. In each row, multiply the f_{xy} by the corresponding coded X , and add along the row. As examples:

(i) In the 4th row: $(1)(2) + (1)(4) = 6$

(iii) In the 7th row: $(1)(-2) + (3)(-1) + (7)(1) + (3)(3) + (3)(4) = 23$

These are entered in the right-hand column, ΣXf_{xy} . Then form the sum of products of this column with the coded Y column, giving $\Sigma XYf_{xy} = 2,318$. The correction term is subtracted as shown in table 7.10.2 to give $\Sigma xy = 2,323.20$.

6. The value of r is now computed (table 7.10.2). No decoding is necessary for r .

As partial checks, the f_x and f_y values both add to the sample size while the column ΣXf_x in step 5 adds to the value ΣXf_x found in step 4.

A large sample provides a good opportunity for checking the assumptions required for the distribution of r . If each number ΣXf_{xy} in the right-hand column is divided by the corresponding f_y , we obtain the mean of X in each array (weight class). These may be plotted against Y to see whether the regression of X on Y appears linear. Similarly, by

TABLE 7 10 1
COMPUTATION OF SAMPLE CORRELATION COEFFICIENT IN TWO WAY FREQUENCY TABLE
(Frequency of occurrence of ears of maize having each diameter and weight)

Y = Weight, Grams	X = Diameter Millimeters																f_y	Coded Y	Yf_y	ΣXf_{xy}
	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51				
320													1				1	12	12	4
310												1					1	11	11	3
300													3				3	10	30	12
290													1				1	9	18	6
280												1			4			8	64	32
270												2		1		1		7	49	23
260							1					3					19	6	114	23
250												2		1			11	5	55	24
240							3					6			1		23	4	92	30
230						1	1					4		2			20	3	60	26
220						4	2					1			1		23	2	46	5
210						1	2					3					28	1	28	4
200				1		3	2					4					29	0	0	7
190						1	2					3		2			32	1	-32	-22
180						2	5					2		1			28	-2	-56	-14
170						4	4					4					21	-3	-63	-20
160			1			2	5					2					19	-4	-76	-15
150			1			2	4					3			1		14	-5	-70	-15
140						3	2					1					11	-6	-66	-22
130						2	1										9	-7	-63	-22
120			1			5	2										3	-8	-24	-33
110						2											3	-9	-27	-4
100																	2	-10	-20	-5
90			1			1											4	-11	-44	-8
80						1	1										3	-12	-36	-9
70																	1	-13	-13	-5
60																	1	-14	-14	-4
50	1																1	-15	-15	-8
f_x	1	0	4	7	18	26	28	51	47	49	31	33	19	4	8	1	327		-46	37
Coded X	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7				
Xf_x	-8	0	-24	-35	-72	-78	-56	-51	0	49	62	99	76	20	48	7	37			

TABLE 7 10 2
CALCULATION OF CORRELATION COEFFICIENT IN TABLE 7 10 1

$\Sigma Xf_x = 37$	$\Sigma Yf_y = -46$	
$\Sigma X^2f_x = 2,279$	$\Sigma Y^2f_y = 7,264$	$\Sigma XYf_{xy} = 2,318$
$(\Sigma Xf_x)^2/n = 419$	$(\Sigma Yf_y)^2/n = 647$	$(\Sigma Xf_x)(\Sigma Yf_y)/n = -520$
$\Sigma x^2 = 2,274.81$	$\Sigma y^2 = 7,257.53$	$\Sigma xy = 2,323.20$
$r = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}} = \frac{2,323.20}{\sqrt{(2,274.81)(7,257.53)}} = 0.5718$		

extra calculation the values ΣYf_{xy} and the Y means may be obtained for each column and plotted against X . A test for linearity of regression is given in section 15.4. The model also assumes that the variances of Y in each column, $s_{y \cdot x}^2$, are estimates of the same quantity, and similarly for the variances of $s_{x \cdot y}^2$ of X within each row. Section 10.21 supplies a test of homogeneity of variances.

EXAMPLE 7 10 1—Using the data in columns f_y and Y table 7 10 1, calculate $\Sigma y^2 = 7,257.53$ together with the sample mean and standard error 198.6 ± 2.61 .

EXAMPLE 7 10 2—Calculate the sample mean, 44.1, and standard deviation, 2.64 in the 42-millimeter array of weights table 7 10 1.

EXAMPLE 7 10 3—In the 200-gram array of diameters compute $\bar{X} = 198.6$ and $s = 47.18$.

EXAMPLE 7 10 4—Compute the sample regression coefficient of weight on diameter 1.0213 together with the regression equation $\hat{Y} = 1.0213X + 154.81$.

EXAMPLE 7 10 5—Calculate the mean diameter in each of the 28 weight arrays. Plot these means against the weight class marks. Does there seem to be any pronounced curvilinearity in the regression of these mean diameters on the weight? Can you write the regression equation giving estimated diameter for each weight?

EXAMPLE 7 10 6—Calculate the sample mean weight of the ears in each of the 16 diameter arrays of table 7 10 1. Present these means graphically as ordinates with the corresponding diameters as abscissas. Plot the graph of the regression equation on the same figure. Do you get a good fit? Is there any evidence of curvilinearity in the regression of means?

7.11—Non-parametric methods. Rank correlation. Often, a bivariate population is far from normal. In that event, the computation of r as an estimate of ρ is no longer valid. In some cases a transformation of the variables X_1 and X_2 brings their joint distribution close to the bivariate normal, making it possible to estimate ρ in the new scale. Failing this, methods of expressing the amount of correlation in non-normal data by means of a parameter like ρ have not proceeded very far.

Nevertheless, we may still want to examine whether two variables are independent or whether they vary in the same or in opposite directions. For a test of the null hypothesis that there is no correlation r may be used provided that one of the variables is normal. When neither variable seems

TABLE 7.11.1
RANKING OF SEVEN RATS BY TWO OBSERVERS OF THEIR CONDITION AFTER THREE WEEKS
ON A DEFICIENT DIET

Rat Number	Ranking by		Difference, d	d^2
	Observer 1	Observer 2		
1	4	4	0	0
2	1	2	-1	1
3	6	5	1	1
4	5	6	-1	1
5	3	1	2	4
6	2	3	-1	1
7	7	7	0	0
			$\Sigma d = 0$	$\Sigma d^2 = 8$

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 8}{7(49 - 1)} = 0.857$$

normal, the best-known procedure is that in which X_1 and X_2 are both rankings. If two judges each rank 12 abstract paintings in order of attractiveness, we may wish to know whether there is any degree of agreement among the rankings. Table 7.11.1 shows similar rankings of the condition of 7 rats after a period of deficient feeding. With data that are not initially ranked, the first step is to rank X_1 and X_2 separately.

The *rank correlation coefficient*, due to Spearman (11) and usually denoted by r_s , is the ordinary correlation coefficient r between the ranked values X_1 and X_2 . It can be calculated in the usual way as $\Sigma(x_1x_2)/\sqrt{(\Sigma x_1^2)(\Sigma x_2^2)}$. An easier method of computing r is given by the formula

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)},$$

whose calculation is explained in table 7.11.1. Like r , the rank correlation can range in samples from -1 (complete discordance) to $+1$ (complete concordance).

For samples of 10 or fewer pairs, the significance levels of r_s , worked out by Kendall (12), (13), are given in table 7.11.2. In the rankings of the rats, $r_s = 0.857$ with 7 pairs. The correlation is significant at the 5% level but not at the 1%. For samples of more than 10 pairs, the null distribution of r_s is similar to that of r , and table A 11 is used for testing r_s . Remember that the degrees of freedom in table A 11 are two less than the number of pairs (size of sample).

Another measure of degree of concordance, closely related to r_s , is Kendall's τ (12). To compute this, rearrange the two rankings so that

TABLE 7.11.2
SIGNIFICANCE LEVELS OF r_s IN SMALL SAMPLES

Size of Sample	5% Level	1% Level
4 or less	none	none
5	1.000	none
6	0.886	1.000
7	0.750	0.893
8	0.714	0.857
9	0.683	0.833
10	0.648	0.794
11 or more	Use table A 11 (p. 557)	

one of them is in the order 1, 2, 3, . . . n . For table 7.11.1, putting observer 1 in this order, we have:

Rat No	2	6	5	1	4	3	7
Observer 1	1	2	3	4	5	6	7
Observer 2	2	3	1	4	6	5	7

Taking each rank given by observer 2 in turn, count how many of the ranks to the *right* of it are smaller than it, and add these counts. For the rank 2 given to rat No. 2 the count is 1, since only rat 5 has a smaller rank. The six counts are 1, 1, 0, 0, 1, 0, there being no need to count the extreme right rank. The total is $Q = 3$. Kendall's τ is

$$\tau = 1 - \frac{4Q}{n(n-1)} = 1 - \frac{12}{42} = \frac{5}{7} = 0.714$$

Like r_s , τ lies between +1 (complete concordance) and -1 (complete disagreement). It takes a little longer to compute, but its frequency distribution on the null hypotheses is simpler and it can be extended to study partial correlation. For details, see (12).

The quantities r_s and τ can be used as a measure of ability to appraise or detect something by ranking. For instance, a group of subjects might each be given bottles containing four different strengths of a delicate perfume and asked to place the bottles in order of the concentration of perfume. If X_1 represents the correct ranking of the strengths and X_2 a subject's ranking, the value of r_s or τ for this subject measures, although rather crudely, his success at this task. From the results for a sample of men and women we could investigate whether women are better at this task than men. The difference between τ of r_s for women and men could be compared, approximately, by an ordinary t -test.

7.12—The comparison of two correlated variances. In section 4.15 (p. 116) we showed how to test the null hypothesis that two *independent*

estimates of variance, s_1^2 and s_2^2 , are each estimates of the same unknown population variance σ^2 . The procedure was to calculate $F = s_1^2/s_2^2$, where s_1^2 is the larger of the two, and refer to table 4.15.1 or table A 14.

This problem arises also when the two estimates s_1^2 and s_2^2 are correlated. For instance, in the sample of pairs of brothers and sisters (section 7.1.), we might wish to test whether brother heights, X_1 , are more or less variable than sister heights, X_2 . We can calculate s_1^2 and s_2^2 , the variances of the two heights between families. But in our sample of 11 families the correlation between X_1 and X_2 was found to be $r = 0.558$. Although this did not reach the 5% level of r (0.602 for 9 d.f.), the presence of a correlation was confirmed by Pearson and Lee's value of $r = 0.553$ for the sample of 1,401 families from which our data were drawn. In another application, a specimen may be sent to two laboratories that make estimates X_1 , X_2 of the concentration of a rare element contained in it. If a number of specimens are sent, we might wish to examine whether one laboratory gives more variability in results than the other.

The test to be described is valid for a sample of pairs of values X_1 , X_2 that follows a bivariate normal. It holds for any value ρ of the population correlation between X_1 and X_2 . If you are confident that ρ is zero, the ordinary F -test should be used, since it is slightly more powerful. When ρ is not zero, the F -test is invalid.

The test is derived by an ingenious approach due to Pitman (15). Suppose that X_1 and X_2 have variances σ_1^2 and σ_2^2 and correlation ρ . The null hypothesis states that $\sigma_1^2 = \sigma_2^2$: for the moment, we are not assuming that the null hypothesis is necessarily true. Since X_1 and X_2 follow a bivariate normal, it is known that $D = X_1 - X_2$ and $S = X_1 + X_2$ also follow a bivariate normal. Let us calculate the correlation ρ_{DS} between D and S . From section 7.9,

$$\begin{aligned}\sigma_D^2 &= \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 \\ \sigma_S^2 &= \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2 \\ \text{Cov.}(DS) &= \text{Cov.}(X_1 - X_2)(X_1 + X_2) = \sigma_1^2 - \sigma_2^2\end{aligned}$$

since the two terms in $\text{Cov.}(X_1X_2)$ cancel. Hence

$$\rho_{DS} = (\sigma_1^2 - \sigma_2^2) / \sqrt{(\sigma_1^2 + \sigma_2^2)^2 - 4\rho^2\sigma_1^2\sigma_2^2}$$

If $\phi = \sigma_1^2/\sigma_2^2$ is the variance-ratio of σ_1^2 to σ_2^2 , this may be written

$$\rho_{DS} = (\phi - 1) / \sqrt{(\phi + 1)^2 - 4\rho^2\phi} \quad (7.12.1)$$

Under the null hypothesis, $\phi = 1$, so that $\rho_{DS} = 0$. If $\sigma_1^2 > \sigma_2^2$, then $\phi > 1$ and ρ_{DS} is positive, while if $\sigma_1^2 < \sigma_2^2$, ρ_{DS} is negative.

Thus, the null hypothesis can be tested by finding D and S for each pair, computing the sample correlation coefficient r_{DS} , and referring to table A 11. A significantly positive value of r_{DS} indicates $\sigma_1^2 > \sigma_2^2$, while a significantly negative one indicates $\sigma_1^2 < \sigma_2^2$.

Alternatively, by the same method that led to equation (7.12.1), r_{DS} can be computed as

$$r_{DS} = (F - 1)/\sqrt{\{(F + 1)^2 - 4r^2F\}}, \quad (7.12.2)$$

where $F = s_1^2/s_2^2$ and r is the correlation between X_1 and X_2 .

In a sample of 173 boys, aged 13–14, height had a standard deviation $s_1 = 5.299$, while leg length gave $s_2 = 4.766$, both figures being expressed as percentages of the sample means (16). The correlation between height and length was $r = 0.878$, a high value, as would be expected. To test whether height is relatively more variable than leg length, we have

$$F = (5.299/4.766)^2 = 1.237$$

and from equation (7.12.2),

$$r_{DS} = (0.237)/\sqrt{\{(2.237)^2 - 4(0.878)^2(1.237)\}} = 0.237/1.136 = 0.209$$

with $d.f. = 173 - 2 = 171$. This value of r_{DS} is significant at the 1% level, since table A 11 gives the 1% level as 0.208 for 150 $d.f.$

The above test is two-tailed: for a one-tailed test, use the 10% and 2% levels in table A 11.

This approach also provides confidence limits for ϕ from a knowledge of F and r . The variates $D' = (X_1/\sigma_1 - X_2/\sigma_2)$ and $S' = (X_1/\sigma_1 + X_2/\sigma_2)$ are uncorrelated whether σ_1 equals σ_2 or not. The sample correlation coefficient between these variates, say R , therefore follows the usual distribution of a sample correlation when $\rho = 0$. As a generalization of formula 7.12.2, the value of R may be shown to be

$$R = (F - \phi)/\sqrt{\{(F + \phi)^2 - 4r^2F\phi\}}$$

In applying this result, it is easier to use the t -table than that of r . The value of t is

$$t = (F - \phi)/\sqrt{n - 2}/2\sqrt{\{(1 - r^2)F\phi\}} \quad (7.12.3)$$

If ϕ is much smaller than F , t becomes large and positive: if ϕ is much larger than F , t becomes large and negative. Values of ϕ that make t lie between the limits $\pm t_{0.05}$ form a 95% confidence interval. The limits found by solving (7.12.3) for ϕ are computed as

$$\phi = F\{K \pm \sqrt{(K^2 - 1)}\},$$

where

$$K = 1 + \frac{2(1 - r^2)t_{0.05}^2}{(n - 2)} \quad d.f. \text{ for } t_{0.05} = n - 2$$

REFERENCES

1. K. PEARSON and A. LEE. *Biometrika*, 2:357 (1902–3).
2. A. E. BRANDT. Ph.D. Thesis, Iowa State University (1932).
3. A. T. CRATHORNE. *Reorganization of Mathematics in Secondary Education*, p. 105, Math. Assoc. of America, Inc. (1923).

- 4 F N DAVID *Tables of the Correlation Coefficient* Cambridge University Press (1938)
- 5 T EDEN *J Agric Sci* 21 547 (1931)
- 6 J M EVVARD M G SNELL, C C CULBERTSON, and G W SNEDECOR *Proc Amer Soc Animal Production* p 2 (1927)
- 7 E S HABER Data from the Iowa Agric Exp Sta
- 8 R A FISHER *Biometrika*, 10 507 (1915)
- 9 R A FISHER *Metron*, 1 3 (1921)
- 10 R A GROUT *Iowa Agric Exp Sta Bul* 218 (1937)
- 11 C SPEARMAN *Amer J Psych*, 15 88 (1904)
- 12 M G KENDALL *Rank Correlation Methods*, 2nd ed, Charles Griffin, London (1955)
- 13 S T DAVID M G KENDALL, and A STUART *Biometrika*, 38 131 (1951)
- 14 J L LUSH *J Agric Res*, 42 853 (1931)
- 15 E J G PITMAN *Biometrika*, 31 9 (1939)
- 16 A A MUMFORD and M YOUNG *Biometrika*, 15 108 (1923)
- 17 C H FISHER *Ann Math Statist* 4 103 (1933)
- 18 F GALTON *Proc Roy Soc London*, 45 135 (1888)
- 19 G UDNY YULE *J Roy Statist Soc*, 89 1 (1926)
- 20 E W LINDSTROM *Amer Nat*, 49 311 (1935)

Sampling from the binomial distribution

8.1—Introduction. In chapter 1 the sampling of attributes was used to introduce some common statistical terms and techniques—estimators, confidence intervals, the binomial distribution, tests of significance, and the chi-square test as applied to a simple proportion. We return to the sampling of attributes in order to fill in the mathematical background of these techniques. The binomial distribution and its relation to the normal distribution will be examined more thoroughly. Further, just as you learned how to compare the means of two normal samples, independent or paired, we shall study the comparison of two proportions from independent samples and from paired samples.

Suppose that an attribute is possessed by a proportion p of the members of a population. A random sample of size n is drawn. The binomial distribution gives a formula for the probability that the sample contains exactly r members having the attribute. The formula is derived from some rules in the theory of probability, now to be explained.

8.2—Some simple rules of probability. The study of probability began around three hundred years ago. At that time, gambling and game of chance had become a fashionable pastime, and there was much interest in questions about the chance that a certain type of card would be drawn from a pack, or that a die would fall in a certain way.

In a problem in probability, we are dealing with a trial, about to be made, that can have a number of different outcomes. A six-sided die, when thrown, may show any of the numbers 1, 2, 3, 4, 5, 6 face upwards—these are the outcomes. Simpler problems in probability can often be solved by writing down all the different possible outcomes of the trial and recognizing that these are *equally likely*. Suppose that the letters a, b, c, d, e, f, g are written on identical balls which are placed in a bag and mixed thoroughly. One ball is drawn out blindly. Most people would say without hesitation that the probability that an a is drawn is $1/7$ because there are 7 balls, one of them is certain to be drawn and all are equally likely. In general terms this result may be stated as follows:

Rule 1. If a trial has k equally likely outcomes, of which one and only one will happen, the probability of any individual outcome is $1/k$.

The claim that the outcomes are equally likely must be justified by knowledge of the exact nature of the trial. For instance, dice to be used in gambling for stakes are manufactured with care to ensure that they are cubes of even density. They are discarded by gambling establishments after a period of use, in case the wear, though not detectable by the naked eye, has made the six outcomes no longer equally likely. The statement that the probability is $1/52$ of drawing the ace of spades from an ordinary pack of cards assumes a thorough shuffling that is difficult to attain, particularly when the cards are at all worn.

In some problems the event in which we are interested will happen if any one of a specific group of outcomes turns up when the trial is made. With the letters a, b, c, d, e, f, g , suppose we ask "what is the probability of drawing a vowel?" The event is now "A vowel is drawn." This will happen if either an a or an e is the outcome. Most people would say that the probability is $2/7$, because there are 2 vowels present out of seven competing letters, and each letter is equally likely. Similarly, the probability that the letter drawn is one of the first four letters is $4/7$. These results are an application of a second rule of probability.

Rule 2. (The Addition Rule). If an event is satisfied by any one of a group of mutually exclusive outcomes, the probability of the event is the sum of the probabilities of the outcomes in the group.

In mathematical terminology, this rule is sometimes stated as:

$$P(E) = P(O_1 \text{ or } O_2 \text{ or } \dots \text{ or } O_m) = P(O_1) + P(O_2) + \dots + P(O_m),$$

where $P(O_i)$ denotes the probability of the i th outcome.

Rule 2 contains one condition: the outcomes in the group must be *mutually exclusive*. This phrase means that if any one of the outcomes happens, all the others fail to happen. The outcomes " a is drawn" and " e is drawn" are mutually exclusive. But the outcomes " a vowel is drawn" and " $one of the first four letters is drawn$ " are not mutually exclusive, because if a vowel is drawn, it might be an a , in which case the event " $one of the first four letters is drawn$ " has also happened.

The condition of mutual exclusiveness is essential. If it does not hold, Rule 2 gives the wrong answer. To illustrate, consider the probability that the letter drawn is either one of the first four letters or is a vowel. Of the seven original outcomes, a, b, c, d, e, f, g , five satisfy the event in question, namely a, b, c, d, e . The probability is given correctly by Rule 2 as $5/7$, because these five outcomes are mutually exclusive. But we might try to shortcut the solution by saying "The probability that one of the first four letters is drawn is $4/7$ and the probability that a vowel is

drawn is $2/7$. Therefore, by Rule 2, the probability that one or the other of these happens is $6/7$." This, you will note, is the wrong answer.

In leading up to the binomial distribution we have to consider the results of repeated drawings from a population. The successive trials or drawings are assumed *independent* of one another. This term means that the outcome of a trial does not depend in any way on what happens in the other trials.

With a series of trials the easier problems can again be solved by Rules 1 and 2. For example, a bag contains the letters a, b, c . In trial 1 a ball is drawn after thorough mixing. The ball is replaced, and in trial 2 a ball is again drawn after thorough mixing. What is the probability that both balls are a ? First, we list all possible outcomes of the two trials. These are $(a, a), (a, b), (a, c), (b, a), (b, b), (b, c), (c, a), (c, b), (c, c)$, where the first letter in a pair is the result of trial 1 and the second that of trial 2. Then we claim that these nine outcomes of the pair of trials are equally likely. Challenged to support this claim, we might say: (i) a, b , and c are equally likely at the first draw, because of the thorough mixing, and, (ii), at the second draw, the conditions of thorough mixing and of independence make all nine outcomes equally likely. The probability of (a, a) is therefore $1/9$.

Similarly, suppose we are asked the probability that the two drawings contain no c 's. This event is satisfied by four mutually exclusive outcomes: $(a, a), (a, b), (b, a)$, and (b, b) . Consequently, the probability (by Rule 2) is $4/9$.

Both the previous results can be obtained more quickly by noticing that the probability of the combined event is the *product* of the probabilities of the desired events in the individual trials. In the first problem the probability of an a is $1/3$ in the first trial and also $1/3$ in the second trial. The probability that both events happen is $1/3 \times 1/3 = 1/9$. In the second problem, the probability of not drawing a c is $2/3$ in each individual trial. The probability of the combined event (no c at either trial) is $2/3 \times 2/3 = 4/9$. A little reflection will show that the numerator of this product (1 or 4) is the number of equally likely outcomes of the two drawings that satisfy the desired combined event. The denominator, 9, is the total number of equally likely outcomes in the combined trials. The probabilities need not be equal at the two drawings. For example, the probability of getting an a at the first trial but not at the second is $1/3 \times 2/3 = 2/9$, the outcomes that produce this event being (a, b) and (a, c) .

Rule 3. (The Multiplication Rule). In a series of independent trials, the probability that each of a specified series of events happens is the product of the probabilities of the individual events.

In mathematical terms,

$$P(E_1 \text{ and } E_2 \dots \text{ and } E_m) = P(E_1)P(E_2) \dots P(E_m)$$

In practice, the assumption that trials are independent, like the assumption that outcomes are equally likely, must be justified by knowledge of the circumstances of the trials. In complex probability problems there have been disputes about the validity of these assumptions in particular applications, and some interesting historical errors have occurred.

This account of probability provides only the minimum background needed for working out the binomial distribution. Reference (1) is recommended as a more thorough introduction to this important subject at an elementary mathematical level.

EXAMPLE 8.2.1—A bag contains the letters $A, b, c, D, e, f, G, h, I$. If each letter is equally likely to be drawn, what is the probability of drawing: (i) a capital letter, (ii) a vowel, (iii) either a capital or a vowel. Ans. (i) $4/9$, (ii) $1/3$, (iii) $5/9$. Does Rule 2 apply to the two events mentioned in (iii)?

EXAMPLE 8.2.2—Three bags contain, respectively, the letters $a, b; c, d, e; f, g, h, i$. A letter is drawn independently from each bag. Write down all 24 equally likely outcomes of the three drawings. Show that six of them give a consonant from each bag. Verify that Rule 3 gives the correct probability of drawing a consonant from each bag ($1/4$).

EXAMPLE 8.2.3—Two six-sided dice are thrown independently. Find the probability: (i) that the first die gives a 6 and the second at least a 3, (ii) that one die gives a 6 and the other at least a 3, (iii) that both give at least a 3, (iv) that the sum of the two scores is not more than 5. Ans. (i) $1/9$, (ii) $2/9$, (iii) $4/9$, (iv) $5/18$.

EXAMPLE 8.2.4—From a bag with the letters a, b, c, d, e a letter is drawn and laid aside, then a second is drawn. By writing down all equally likely pairs of outcomes, show that the probability that both letters are vowels is $1/10$. This is a problem to which Rule 3 does not apply. Why not?

EXAMPLE 8.2.5—If two trials are not independent, the probability that event E_1 happens at the first trial and E_2 at the second is obtained (1) by a generalization of Rule 3: $P(E_1 \text{ and } E_2) = P(E_1)P(E_2, \text{ given that } E_1 \text{ has happened})$. This last factor is called the *conditional probability of E_2 given E_1* , and is usually written $P(E_2|E_1)$. Show that this rule gives the answer, $1/10$, in example 8.2.4, where E_1, E_2 are the probabilities of drawing a vowel at the first and second trials, respectively.

In many applications, the probability of a particular outcome must be determined by a statistical study. For instance, insurance companies are interested in the probability that a man aged sixty will live for the next ten years. This quantity is calculated from national statistics of the age distribution of males and of the age distribution of deaths of males, and is published in actuarial tables. Provided that the conditions of independence and of mutually exclusive outcomes hold where necessary, Rules 2 and 3 are applied to probabilities of this type also. Thus, the probability that three men aged sixty, selected at random from a population, will all survive for ten years would be taken as p^3 , where p is the probability that an individual sixty-year-old man will survive for ten years.

8.3—The binomial distribution. A proportion p of the members of a population possess some attribute. A sample of size $n = 2$ is drawn. The result of a trial is denoted by S (success) if the member drawn has the attribute and by F (failure) if it does not. In a single drawing, p is the

TABLE 8.3.1
THE BINOMIAL DISTRIBUTION FOR $n = 2$

(1) Outcomes of Trial		(2) Probability	(3) No. of Successes	(4) Probability
1	2			
<i>F</i>	<i>F</i>	qq	0	q^2
<i>F</i>	<i>S</i>	qp	1	$2pq$
<i>S</i>	<i>F</i>	pq		
<i>S</i>	<i>S</i>	pp	2	p^2
Total		1		1

probability of obtaining an *S*, while $q = 1 - p$ is the probability of obtaining an *F*. Table 8.3.1 shows the four mutually exclusive outcomes of the two drawings, in terms of successes and failures.

The probabilities given in column (2) are obtained by applying Rule 3 to the two trials. For example, the probability of two successive *F*'s is qq , or q^2 . This assumes, of course, that the two trials are independent, as is necessary if the binomial distribution is to hold. Coming to the third column, we are now counting the number of successes. Since the two middle outcomes, *FS* and *SF*, both give 1 success, the probability of 1 success is $2pq$ by Rule 2. The third and fourth columns present the binomial distribution for $n = 2$. As a check, the probabilities in columns 2 and 4 each add to unity, since

$$q^2 + 2pq + p^2 = (q + p)^2 = (1)^2 = 1$$

TABLE 8.3.2
THE BINOMIAL DISTRIBUTION FOR $n = 3$

(1) Outcomes of Trial			(2) Probability	(3) No. of Successes	(4) Probability
1	2	3			
<i>F</i>	<i>F</i>	<i>F</i>	qqq	0	q^3
<i>F</i>	<i>F</i>	<i>S</i>	qqp	1	$3pq^2$
<i>F</i>	<i>S</i>	<i>F</i>	qpq		
<i>S</i>	<i>F</i>	<i>F</i>	pqq		
<i>F</i>	<i>S</i>	<i>S</i>	qpp	2	$3p^2q$
<i>S</i>	<i>F</i>	<i>S</i>	pqp		
<i>S</i>	<i>S</i>	<i>F</i>	ppq		
<i>S</i>	<i>S</i>	<i>S</i>	ppp	3	p^3

In the same way, table 8.3.2 lists the eight relevant outcomes for $n = 3$. The probabilities in the second and fourth columns are obtained by Rules 3 and 2 as before. Three outcomes provide 1 success, with total probability $3pq^2$, while three provide 2 successes with total probability $3p^2q$. Check that the eight outcomes in the first column are mutually exclusive.

The general structure of the binomial formula is now apparent. The formula for the probability of r successes in n trials has two parts. One part is the term $p^r q^{n-r}$. This follows from Rule 3, since any outcome of this type must have r S 's and $(n - r)$ F 's in the set of n draws. The other part is the number of mutually exclusive ways in which the r S 's and the $(n - r)$ F 's can be arranged. In algebra this term is called the number of combinations of r letters out of n letters. It is denoted by the symbol $\binom{n}{r}$. The formula is

$$\binom{n}{r} = \frac{n(n - 1)(n - 2) \dots (n - r + 1)}{r(r - 1)(r - 2) \dots (2)(1)}$$

For small samples these quantities, the *binomial coefficients*, can be written down by an old device known as *Pascal's triangle*, shown in table 8.3.3.

Each coefficient is the sum of the two just above it to the right and the left. Thus, for $n = 8$, the number $56 = 21 + 35$. Note that for any n the coefficients are symmetrical, rising to a peak in the middle.

Putting the two parts together, the probability of r successes in a sample of size n is

$$\binom{n}{r} p^r q^{n-r} = \frac{n(n - 1)(n - 2) \dots (n - r + 1)}{r(r - 1)(r - 2) \dots (2)(1)} p^r q^{n-r}$$

These probabilities are the successive terms in the expansion of the binomial expression $(q + p)^n$. This fact explains why the distribution is called binomial, and also verifies that the sum of the probabilities is 1, since $(q + p)^n = (1)^n = 1$.

TABLE 8.3.3
BINOMIAL COEFFICIENTS GIVEN BY PASCAL'S TRIANGLE

Size of Sample	Binomial Coefficients									
n										
1						1				
2					1	2	1			
3				1	3	3	1			
4			1	4	6	4	1			
5			1	5	10	10	5	1		
6		1	6	15	20	15	6	1		
7		1	7	21	35	35	21	7	1	
8	1	8	28	56	70	56	28	8	1	
					etc.					

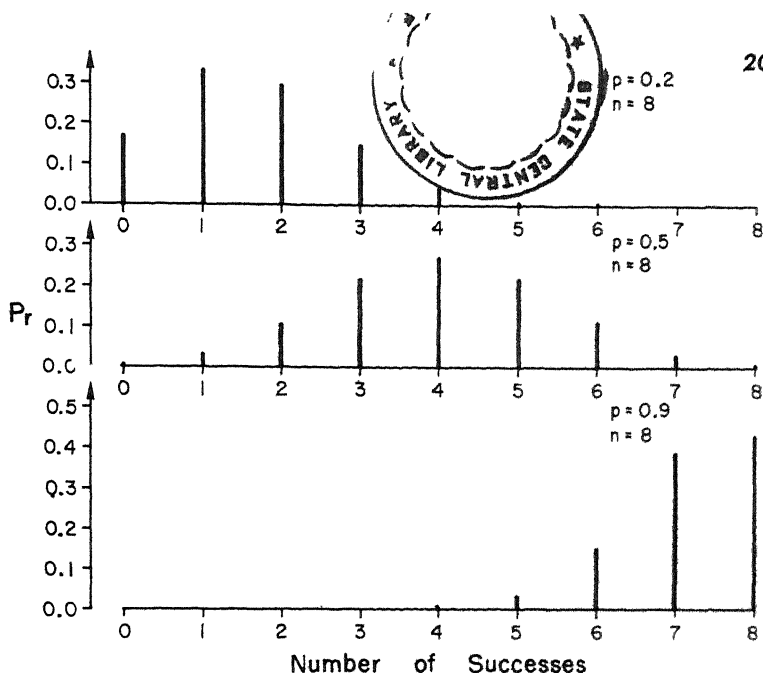


FIG. 8.3.1—Binomial distributions for $n = 8$. Top: $p = 0.2$, Middle: $p = 0.5$, Bottom: $p = 0.9$.

For $n = 8$, figure 8.3.1 shows these distributions for $p = 0.2, 0.5$, and 0.9 . The distribution is positively skew for p less than 0.5 and negatively skew for p greater than 0.5 . For $p = 0.5$ the general shape, despite the discreteness, bears some resemblance to a normal distribution.

Reference (16) contains extensive tables of individual and cumulative terms of the binomial distribution for n up to 49; reference (17) has cumulative terms up to $n = 1,000$.

8.4—Sampling the binomial distribution. As usual, you will find it instructive to verify the preceding theory by sampling. The table of random digits (table A 1, p. 543) is very convenient for drawing samples from the binomial with $n = 5$, since the digits in a row are arranged in groups of 5. For instance, to sample the binomial with $p = 0.2$, let the digits 0 and 1 represent a success, and all other digits a failure. By recording the total number of 0's and 1's in each group of 5, many samples from $n = 5, p = 0.2$ can be drawn quickly. Table 8.4.1 shows the results of 100 drawings of this type, and illustrates a common method of tallying the results. A slanting line is used at every fifth tally, so that $| \uparrow \uparrow \uparrow$ represents 5 drawings of a particular number of successes.

To fit the corresponding theoretical distribution, first calculate the terms $p^r q^{n-r}$. For $r = 0$ (no successes) this is $q^n = (0.8)^5 = 0.32768$. For $r = 1$, it is $pq^{n-1} = (0.2)(0.8)^4$. To obtain a shortcut, notice that this term

TABLE 8.4.1
TALLYING OF 100 DRAWINGS FROM THE BINOMIAL WITH $n = 5, p = 0.2$

No. of Successes		Total
0		32
1		44
2		17
3		6
4		1
5		0
		100

can be written: $(q^n)(p/q)$. It is computed from the previous term by multiplying by $p/q = 0.2/0.8 = 1/4$. Thus for $r = 1$ the term is $(0.32768)/4 = 0.08192$. Similarly, the term for $r = 2$, p^2q^{n-2} , is found by multiplying the term for $r = 1$ by (p/q) , and so on for each successive term. The details appear in table 8.4.2. The binomial coefficients are read from Pascal's triangle. These coefficients and the terms in $p^r q^{n-r}$ are multiplied to give the theoretical probabilities of 0, 1, 2, . . . 5 successes. Finally, since $N = 100$ samples were drawn, we multiply each probability by 100 to give the expected frequencies of 0, 1, 2, . . . 5 successes.

TABLE 8.4.2
FITTING THE THEORETICAL BINOMIAL FOR $n = 5, p = 0.2$

No of Successes (r)	Term $p^r q^{n-r}$	Binomial Coefficient	$\binom{n}{r} p^r q^{n-r}$	Expected Frequency	Observed Frequency
0	0.32768	1	0.32768	32.77	32
1	0.08192	5	0.40960	40.96	44
2	0.02048	10	0.20480	20.48	17
3	0.00512	10	0.05120	5.12	6
4	0.00128	5	0.00640	0.64	1
5	0.00032	1	0.00032	0.03	0
			1.00000	100.00	100

Because of sampling variation, the expected and observed frequencies do not agree exactly, but their closeness is reassuring. Later (section 9.4) a method is given for testing whether the observed and expected frequencies differ by no more than is usual from sampling variation. In the present example, the agreement is in fact better than is usually found in such sampling experiments (example 9.4.1).

EXAMPLE 8.4.1—With $n = 2, p = 1/2$, show that the probability of one success is $1/2$. If p differs from $1/2$, does the probability of one success increase or decrease?

EXAMPLE 8.4.2—A railway company claims that 95% of its trains arrive on time. If a man travels on three of these trains, what is the probability that: (i) all three arrive on time, (ii) one of the three is late, assuming that the claim is correct. Ans (i) 0.857, (ii) 0.135.

EXAMPLE 8.4.3—Assuming that the probability that a child is male is $1/2$, find the probability that in a family of 6 children there are: (i) no boys, (ii) exactly 3 boys, (iii) at least 2 girls, (iv) at least one girl and 1 boy. Ans. (i) $1/64$, (ii) $5/16$, (iii) $57/64$, (iv) $31/32$.

EXAMPLE 8.4.4—Work out the terms of the binomial distribution for $n = 4$, $p = 0.4$. Verify that: (i) the sum of the terms is unity, (ii) 1 and 2 successes are equally probable, (iii) 0 successes is about five times as probable as 4 successes

EXAMPLE 8.4.5—By extending Pascal's triangle, obtain the binomial coefficients for $n = 10$. Hence compute and graph the binomial distribution for $n = 10$, $p = 1/2$. Does the shape appear similar to a normal distribution? Hint: when $p = 1/2$, the term $p^r q^{n-r} = 1/2^n$ for any r . Since $2^{10} = 1,024 \approx 1,000$, the distribution is given accurately enough for graphing by simply dividing the binomial coefficients by 1,000.

8.5—Mean and standard deviation of the binomial distribution. If

$$f_r = \frac{n(n-1) \dots (n-r+1)}{r(r-1) \dots (2)(1)} p^r q^{n-r}$$

denotes the binomial probability of r successes in a sample of size n , the mean and variance of the distribution of the number of successes r are defined by the equations

$$\mu = \sum_{r=0}^n r f_r \quad \sigma^2 = \sum_{r=0}^n (r - \mu)^2 f_r \quad (8.5.1)$$

Note the formula for σ^2 . In a theoretical distribution, σ^2 is the average value of the squared deviation from the population mean. Each squared deviation, $(r - \mu)^2$, is multiplied by its relative frequency of occurrence f_r . The concept of number of degrees of freedom does not come in.

By algebra, it is found from (8.5.1) that

$$\mu = np \quad \sigma^2 = npq \quad \sigma = \sqrt{npq} \quad (8.5.2)$$

These results apply to the *number* of successes. Often, interest centers in the *proportion* of successes, r/n . For this,

$$\mu = p \quad \sigma^2 = pq/n \quad \sigma = \sqrt{pq/n} \quad (8.5.3)$$

Sometimes results are presented in terms of the *percentage* of successes $100r/n$. Formulas (8.5.3) also hold for the percentage of successes if p now stands for the percentage in the population and $q = 100 - p$.

As illustrations, the formulas work out as follows for $n = 64$, $p = 0.2$:

$$\begin{aligned} \text{Number: } \mu &= (64)(0.2) = 12.8 & \sigma &= \sqrt{\{(64)(0.2)(0.8)\}} = \sqrt{10.24} = 3.2 \\ \text{Proportion: } \mu &= 0.2 & \sigma &= \sqrt{\{(0.2)(0.8)/64\}} = \sqrt{0.0025} = 0.05 \\ \text{Percentage: } \mu &= 20 & \sigma &= \sqrt{\{(20)(80)/64\}} = \sqrt{25} = 5 \end{aligned}$$

For a sample of fixed size n , the standard deviations \sqrt{npq} for the number of successes and $\sqrt{pq/n}$ for the proportion of successes are greatest when $p = 1/2$. As p moves towards either 0 or 1, the standard deviation declines, though quite slowly at first, as the following table of \sqrt{pq} shows.

p	0.5	0.4 or 0.6	0.6 or 0.7	0.2 or 0.8	0.1 or 0.9
\sqrt{pq}	0.500	0.490	0.458	0.400	0.300

EXAMPLE 8.5.1—For the binomial distribution of the number of successes with $n = 2$ (given in table 8.3.1, p. 203), verify from formulas 8.5.1 that $\mu = 2p$, $\sigma^2 = 2pq$.

EXAMPLE 8.5.2—For the binomial distribution with $n = 5$, $p = 0.2$, given in table 8.4.2, compute Σrf_r and $\Sigma(r - \mu)^2 f_r$, and verify that the results are $\mu = 1$ and $\sigma^2 = 0.80$.

EXAMPLE 8.5.3—For $n = 96$, $p = 0.4$, calculate the S.D.'s of: (i) the number, (ii) the percentage of successes. Ans. (i) 4.8, (ii) 5.

EXAMPLE 8.5.4—An investigator intends to estimate, by random sampling from a large file of house records, the percentage of houses in a town that have been sold in the last year. He thinks that p is about 10% and would like the standard deviation of his estimated percentage to be about 1%. How large should n be? Ans. 900 houses

There is an easy way of obtaining the results $\mu = p$ and $\sigma^2 = pq/n$ for the distribution of the proportion of successes r/n in a sample of size n . Attach the number 1 to every success in the population and the number 0 to every failure. Instead of thinking of the population as a large collection of the letters S and F , we think of it as a large collection of 1's and 0's. It is the population distribution of a variable X that takes only two values: 1 with relative frequency p and 0 with relative frequency q . The population mean and variance of the new variate X are easily found by working out the definitions (8.5.1),

$$\mu_X = \Sigma X f_X \qquad \sigma_X^2 = \Sigma (X - \mu)^2 f_X$$

where the sum extends only over the two values $X = 0$ and $X = 1$, as shown below:

X	f_X	$X f_X$	$X - \mu$	$(X - \mu)^2$	$(X - \mu)^2 f_X$
0	q	0	$-p$	p^2	$p^2 q$
1	p	p	$1 - p$	q^2	$q^2 p$
$\mu_X = p$			$\sigma_X^2 = pq$		

The variate X has population mean p and population variance pq .

Now draw a random sample of size n . If the sample contains r successes, then ΣX , taken over the sample, is r , so that $\bar{X} = \Sigma X/n$ is r/n , the sample proportion of successes. But we know that the mean of a random sample from any distribution is an unbiased estimate of the population mean, and has variance σ^2/n (section 2.11). Hence $\bar{X} = r/n$ is an unbiased estimate of p , with variance $\sigma_X^2/n = pq/n$.

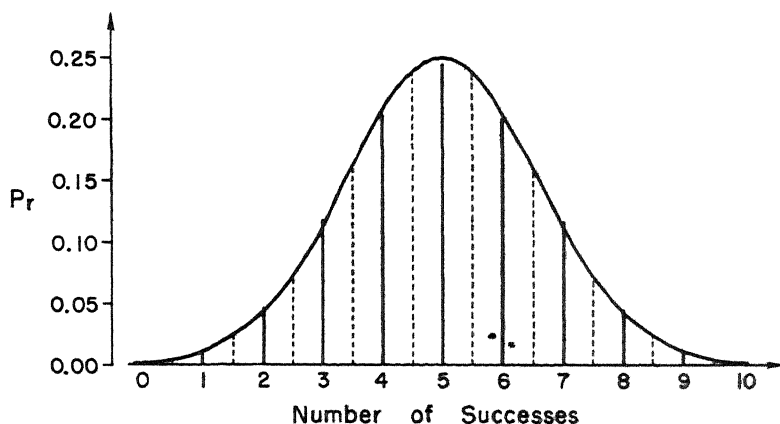


FIG 8.6.1—The solid vertical lines show the binomial distribution of the number of successes for $n = 10$, $p = 0.5$. The curve is the normal approximation to this distribution, which has mean $np = 5$ and $S.D. \sqrt{npq} = 1.581$.

Further, since $X = r/n$ is the mean of a sample from a population that has a finite variance pq , we can quote the Central Limit Theorem (section 2.12). This states that the mean \bar{X} of a random sample from any population with finite variance tends to normality. Hence, as n increases, the binomial distribution of r/n or of r approaches the normal distribution.

For $p = 0.5$ the normal is a good approximation when n is as low as 10. As p approaches 0 or 1, some skewness remains in the binomial distribution until n is large.

8.6—The normal approximation and the correction for continuity. The solid vertical lines in figure 8.6.1 show the binomial distribution of r for $n = 10$, $p = 0.5$. Also shown is the approximating normal curve, with mean $np = 5$ and $S.D. \sqrt{npq} = 1.581$. The normal seems a good approximation to the shape of the binomial.

One difference, however, is that the binomial is discrete, having probability only at the values $r = 0, 1, 2, \dots, 10$, while the normal has probability in any interval from $-\infty$ to ∞ . This raises a problem: in estimating the binomial probability of, say, 4 successes, what part of the normal curve do we use as an approximation? We need to set up a correspondence between the set of binomial ordinates and the areas under the normal curve.

The simplest way of doing this is to regard the binomial as a grouping of the normal into unit class intervals. Under this rule the binomial ordinate at 4 corresponds to the area under the normal curve from $3\frac{1}{2}$ to $4\frac{1}{2}$. The ordinate at 5 corresponds to the area from $4\frac{1}{2}$ to $5\frac{1}{2}$, and so on. The ordinate at 10 corresponds to the normal area from $9\frac{1}{2}$ to ∞ . These class boundaries are the dotted lines in figure 8.6.1.

In the commonest binomial problems we wish to calculate the prob-

abilities at the ends of the distribution; for instance, the probability of 8 or more successes. The exact result, found by adding the binomial probabilities for $r = 8, 9, 10$, is $56/1024 = 0.0547$. Under our rule, the corresponding area under the normal curve is the area from $7\frac{1}{2}$ to ∞ , *not* the area from 8 to ∞ . The normal deviate is therefore $z = (7.5 - 5)/1.581$, which by a coincidence is also 1.581. The approximate probability from the normal table is $P = 0.0570$, close enough to 0.0547. Use of $z = (8 - 5)/1.581$ gives $P = 0.0288$, a poor result.

Similarly, the probability of 4 or fewer successes is approximated by the area of the normal curve from $-\infty$ to $4\frac{1}{2}$. The general rule is to decrease the *absolute* value of $(r - np)$ by $\frac{1}{2}$. Thus,

$$z_c = (|r - np| - \frac{1}{2})/\sqrt{npq}$$

The subtraction of $\frac{1}{2}$ is called the *correction for continuity*. It is simple to apply and usually improves the accuracy of the normal approximation, although when n is large it has only a minor effect.

If you are working in terms of r/n instead of r , then

$$z_c = \frac{|r/n - p| - 1/2n}{\sqrt{(pq/n)}}$$

EXAMPLE 8.6.1—For $n = 10$, $p = 1/2$, calculate: (i) the exact probability of 4 or fewer successes, and the normal approximation, (ii) corrected for continuity, (iii) uncorrected. Ans. (i) 0.377, (ii) 0.376, (iii) 0.263.

EXAMPLE 8.6.2—In a sample of size 49 with $p = 0.2$, the expected number of successes is 9.8. An investigator is interested in the probability that the observed number of successes will be (i) 15 or more, or (ii) 5 or less. Estimate these two probabilities by the corrected normal approximation. Ans. (i) 0.0466 (ii) 0.0623. The exact answers by summing the binomial are: (i) 0.0517, (ii) 0.0547. Because of the skewness ($p = 0.2$), the normal curve underestimates in the long tail and overestimates in the short tail. For the sum of the two tails the normal curve does better, giving 0.1089 as against the exact 0.1064.

EXAMPLE 8.6.3 With $n = 16$, $p = 0.9$, estimate by the normal curve the probability that 16 successes are obtained. The exact result is, of course, $(0.9)^{16} = 0.185$. Ans. 0.180.

8.7—Confidence limits for a proportion. If r members out of a sample of size n are found to possess some attribute, the sample estimate of the proportion in the population possessing this attribute is $\hat{p} = r/n$. In large samples, as we have seen, the binomial estimate \hat{p} is approximately normally distributed about the population proportion p with standard deviation $\sqrt{(pq/n)}$. For the true but unknown standard deviation $\sqrt{(pq/n)}$ we substitute the sample estimate $\sqrt{(\hat{p}\hat{q}/n)}$. Hence, the probability is approximately 0.95 that \hat{p} lies between the limits

$$p - 1.96\sqrt{(\hat{p}\hat{q}/n)} \text{ and } p + 1.96\sqrt{(\hat{p}\hat{q}/n)}$$

But this statement is equivalent to saying that p lies between

$$\hat{p} - 1.96\sqrt{(\hat{p}\hat{q}/n)} \text{ and } \hat{p} + 1.96\sqrt{(\hat{p}\hat{q}/n)} \quad (8.7.1)$$

unless we were unfortunate in drawing one of the extreme samples that

is up once in twenty times. The limits 8.7.1 are therefore the approximate 95% confidence limits for p .

For example, suppose that 200 individuals in a sample of 1,000 possess the attribute. The 95% confidence limits for p are

$$0.2 \pm 1.96\sqrt{(0.2)(0.8)/1000} = 0.2 \pm 0.025$$

The confidence interval extends from 0.175 to 0.225; that is, from 17.5% to 22.5%. Limits corresponding to other confidence probabilities are of course obtained by inserting the appropriate values of the normal deviate. For 99% limits, we replace 1.96 by 2.576.

If the above reasoning is repeated with the correction for continuity included, the 95% limits for p become

$$\hat{p} \pm \{1.96\sqrt{(\hat{p}\hat{q}/n) + 1/2n}\}$$

The correction is easily applied. It amounts to widening the limits a little. We recommend that the correction be used as a standard practice, although it makes little difference when n is large. To illustrate the correction on a smaller sample, suppose that 10 families out of 50 report ownership of more than one car, giving $\hat{p} = 0.2$. The 95% confidence limits for p are

$$0.2 \pm \{1.96\sqrt{0.16/50 + 0.01}\} = 0.2 \pm 0.12,$$

or .08 and .32. More exact limits for this problem, computed from the binomial distribution itself, were presented in table 1.4.1 (p. 6) as 0.10 and 0.3. The normal approximation gives the correct width of the interval, .24, but the normal limits are symmetrical about \hat{p} , whereas the correct limits are displaced upwards because an appreciable amount of skewness still remains in the binomial when $n = 50$ and p is not near 1/2.

If you prefer to express \hat{p} and p in percentages, the 95% limits are

$$\hat{p} \pm \{1.96\sqrt{\hat{p}(100 - \hat{p})/n + 50/n}\}$$

You may verify that this formula gives 8% and 32% as the limits in the above problem.

8.8—Test of significance of a binomial proportion. The normal approximations useful also in testing the null hypothesis that the population proportion of successes has a known value p . If the null hypothesis is true, \hat{p} is distributed approximately normally with mean p and $S.D.$ $\sqrt{(pq/n)}$. With the correction for continuity, the normal deviate is

$$\begin{aligned} z_c &= (|\hat{p} - p| - 1/2n)/\sqrt{(pq/n)} \\ &= (|r - np| - \frac{1}{2})/\sqrt{(npq)} \end{aligned}$$

This can be referred to the normal tables to compute the probability of getting a sample proportion as divergent as the observed one.

To take an example considered in chapter 1, a physician found 480 men and 420 women among 900 admitted to a hospital with a certain

disease. Is this result consistent with the hypothesis that in the population of hospital patients, half the cases are male? Taking r as the number of males,

$$z_c = \frac{|480 - 450| - \frac{1}{2}}{\sqrt{\{(900)(\frac{1}{2})(\frac{1}{2})\}}} = \frac{29.5}{15} = 1.967$$

Since the probability is just on the 5% level, the null hypothesis is rejected at this level.

If the alternative hypothesis is one-tailed, for instance that more than half the hospital patients are male, only one tail of the normal distribution is used. For this alternative the null hypothesis in the example is rejected at the $2\frac{1}{2}\%$ level.

In sections 1.10–1.12 you were given another method of testing a null hypothesis about p by means of chi-square with 1 degree of freedom. In the notation of chapter 1,

$$\chi^2 = \sum \frac{(\text{Obs.} - \text{Exp.})^2}{\text{Exp.}} = \sum \frac{(f - F)^2}{F},$$

the sum being taken over the two classes, male and female. The χ^2 test is exactly the same as the two-tailed z test, except that the above formula for χ^2 contains no correction for continuity. To show the relationship, we need to translate the notation of chapter 1 into the present notation, as follows:

Notation of Chapter 1		Present Notation	
		Males	Females
Observed nos.	f	r	$n - r$
Expected nos	F	np	$nq = n - np$
Obs - Exp	$f - F$	$r - np$	$-(r - np)$

Hence,

$$\begin{aligned} \chi^2 &= \sum \frac{(f - F)^2}{F} = \frac{(r - np)^2}{np} + \frac{(r - np)^2}{nq} \\ &= \frac{(r - np)^2}{npq} (q + p) = \frac{(r - np)^2}{npq} = z^2, \end{aligned}$$

since the normal deviate $z = (r - np)/\sqrt{(npq)}$ if no correction for continuity is used. Further, the χ^2 distribution, with 1 d.f., is the distribution of the square of a normal deviate: the 5% significance level of χ^2 , 3.84, is simply the square of 1.96. Thus, the two tests are identical.

To correct χ^2 for continuity, we use the square of z , corrected for continuity.

$$\chi_c^2 = \frac{(|r - np| - \frac{1}{2})^2}{npq}$$

As with z , we recommend that the correction be applied routinely. For one-sided alternatives the z method is preferable, since χ^2 takes no account of the sign of $(r - np)$ and is basically two-sided.

EXAMPLE 8.8.1—Two workers A and B perform a task in which carelessness leads to minor accidents. In the first 20 accidents, 13 happened to A and 7 to B . In a previous example (1.15.1) you were asked to calculate χ^2 for testing the null hypothesis that A and B are equally likely to have accidents, the answer being $\chi^2 = 1.8$, with P about 0.18. Recalculate χ^2 and P , corrected for continuity. Ans $\chi_c^2 = 1.25$, P slightly greater than 0.25.

EXAMPLE 8.8.2—A question that is asked occasionally is whether the $1/2$ correction should be applied in χ^2 if $|r - np|$ is less than $1/2$. This happens for instance, if $r = 6$, $n = 25$ and the null hypothesis is $p = 1/4$, because $np = 6.25$ and $|r - np| = 0.25$. Strictly, the answer in such cases is that the corrected value of χ^2 is zero. When $n = 25$, the result $r = 6$ is the sample result that gives the closest possible agreement with the null hypothesis, $np = 6.25$. Hence, all possible samples with $n = 25$ give results at least as divergent from the null hypothesis. The significance P is therefore 1, corresponding to $\chi^2 = 0$.

8.9—The comparison of proportions in paired samples. A comparison of two sample proportions may arise either in paired or in independent samples. To illustrate paired samples, suppose that a lecture method is being compared with a method that uses a machine for programmed learning but no lecture, the objective being to teach workers how to perform a rather complicated operation. The workers are first grouped into pairs by means of an initial estimate of their aptitudes for this kind of task. One member of each pair is assigned at random to each method. At the end, each student is tested to see whether he succeeds or fails in a test on the operation.

With 100 pairs, the results might be presented as follows:

Result for Method		No. of Pairs
<i>A</i>	<i>B</i>	
<i>S</i>	<i>S</i>	52
<i>S</i>	<i>F</i>	21
<i>F</i>	<i>S</i>	9
<i>F</i>	<i>F</i>	18
Total		100

In 52 pairs, both workers succeeded in the test; in 21 pairs, the worker taught by method A succeeded, but his partner taught by method B failed, and so on.

As a second illustration (2), different media for growing diphtheria bacilli were compared. Swabs were taken from the throats of a large number of patients with symptoms suggestive of the presence of diphtheria bacilli. From each swab, a sample was grown on each medium. After allowing time for growth, each culture was examined for the presence or

absence of the bacilli. A successful medium is one favorable to the growth of the bacilli so that they are detected. This is an example of self-pairing, since each medium is tested on every patient. It is also an example in which a large number of *FF*'s would be expected, because diphtheria is now rare and many patients would actually have no diphtheria bacilli in their throats.

Consider first the test of significance of the null hypothesis that the proportion of successes is the same for the two methods or media. The *SS* and *FF* pairs are ignored in the test of significance, since they give no indication in favor of either *A* or *B*. We concentrate on the *SF* and *FS* pairs. If the null hypothesis is true, the population must contain as many *SF* as *FS* pairs. In the numerical example there are $21 + 9 = 30$ pairs of the *SF* or *FS* types. Under the null hypothesis we expect 15 of each type as against 21 and 9 observed.

Hence, the null hypothesis is tested by either the χ^2 or the *z* test of the preceding section. (In the *z* test we take $n = 30$, $r = 21$, $p = 1/2$). When $p = 1/2$, χ^2 takes the particularly simple form (section 5.4),

$$\chi_c^2 = \frac{(|21 - 9| - 1)^2}{30} = \frac{121}{30} = 4.03$$

with 1 *d.f.* The null hypothesis is rejected at the 5% level (3.84). Method *A* has given a significantly higher proportion of successes. Remember that in this test, the denominator of χ^2 is always the total number of *SF* and *FS* pairs. This test is the same as the sign test (section 5.4).

The investigator will also be interested in the actual percentages of successes given by the two methods. These were: $52 + 21 = 73\%$ for *A* and $52 + 9 = 61\%$ for *B*. If the task is exceptionally difficult, he might conclude that although *A* is significantly better than *B*, both methods are successful enough to be useful. In other circumstances, he might report that neither method is satisfactory. This might be the case if *A* and *B* were two new techniques for predicting some feature of the weather, and if standard techniques were known to give more than 85% successes.

When there is clearly a difference between the performances of the two methods, we may wish to report this difference, $(73\% - 61\%) = 12\%$, along with its standard error. Let

$$p_{SF} = \text{proportion of } SF \text{ pairs} = \frac{21}{100} = 0.21$$

$$p_{FS} = \text{proportion of } FS \text{ pairs} = \frac{9}{100} = 0.09$$

When the difference is expressed in percentages (12%), a simple formula for its standard error is

$$\begin{aligned}
 S.E. &= 100 \sqrt{\left\{ \frac{p_{SF} + p_{FS} - (p_{SF} - p_{FS})^2}{n} \right\}} \\
 &= 100 \sqrt{\left\{ \frac{0.21 + 0.09 - (0.21 - 0.09)^2}{100} \right\}} \\
 &= 10\sqrt{0.2856} = 5.3
 \end{aligned}$$

If the difference is expressed in proportions, the factor 100 is omitted.

Note: If you record only that *A* gave 73 successes and *B* gave 61 successes out of 100, the test of significance in paired data cannot be made from this information alone. The classification of the results for the individual pairs must be available.

8.10—Comparison of proportions in two independent samples: the 2×2 table. This problem occurs very often in investigative work. Many controlled experiments which compare two procedures or treatments are carried out with independent samples, because no effective way of pairing the subjects or animals is known to the investigator. Comparison of proportions in different groups is also common in non-experimental studies. A manufacturer compares the proportions of defective articles found in two separate sources of supply from which he buys these articles, or a safety engineer compares the proportions of head injuries sustained in automobile accidents by passengers with seat belts and those without seat belts.

Alternatively, a single sample may be classified according to two different attributes. The data used to illustrate the calculations come from a large Canadian study (3) of the relation between smoking and mortality. By an initial questionnaire in 1956, male recipients of war pensions were classified according to their smoking habits. We shall consider two classes: (i) non-smokers and (ii) those who reported that they smoked pipes only. For any pensioner who died during the succeeding six years, a report of the death was obtained. Thus, the pensioners were classified also according to their status (dead or alive) at the end of six years. Since the probability of dying depends greatly on age, the comparison given here is confined to men aged 60–64 at the beginning of the study. The numbers of men falling in the four classes are given in table 8.10.1, called a 2×2 contingency table.

It will be noted that 11.0% of the non-smokers had died, as against 13.4% of the pipe smokers. Can this difference be attributed to sampling error, or does it indicate a real difference in the death rates in the two groups? The null hypothesis is that the proportions dead, 117/1067 and 54/402, are estimates of the same quantity.

The test can be performed by χ^2 . As usual,

$$\chi^2 = \sum \frac{(f - F)^2}{F},$$

TABLE 8.10.1
MEN CLASSIFIED BY SMOKING HABIT AND MORTALITY IN SIX YEARS

	Non-smokers	Pipe Smokers	Total
Dead	117	54	171
Alive	950	348	1,298
Total	1,067	402	1,469
% dead	11.0	13.4	

where the f 's are the observed numbers 117, 950, 54, 348 in the four cells. The F 's are the numbers that would be expected in the four cells if the null hypothesis were true.

The F 's are computed as follows. If the proportions dead are the same for the two smoking classes, our best estimate of this proportion is the proportion, $171/1469$, found in the combined sample. Since there are 1067 non-smokers, the expected number dead, on the null hypothesis, is

$$\frac{(1067)(171)}{1469} = 124.2$$

The rule is: to find the expected number in any cell, multiply the corresponding column and row totals and divide by the grand total. The expected number of non-smokers who are alive is

$$\frac{(1067)(1298)}{1469} = 942.8,$$

and so on. Alternatively, having calculated 124.2 as the expected number of non-smokers who are dead, the expected number alive is found more easily as $1067 - 124.2 = 942.8$. Similarly, the expected number of pipe smokers who are dead is $171 - 124.2 = 46.8$. Finally, the expected number of pipe smokers who are alive is $402 - 46.8 = 355.2$. Thus, only *one* expected number need be calculated; the others are found by subtraction. The observed numbers, expected numbers, and the differences ($f - F$) appear in table 8.10.2.

Except for their signs, all four deviations ($f - F$) are equal. This result holds in any 2×2 table.

TABLE 8.10.2
VALUES OF f (OBSERVED), F (EXPECTED), AND $(f - F)$ IN THE FOUR CELLS

f		F		$f - F$	
117	54	124.2	46.8	-7.2	+7.2
950	348	942.8	355.2	+7.2	-7.2

Since $(f - F)^2$ is the same in all cells, χ^2 may be written

$$\begin{aligned}\chi^2 &= (f - F)^2 \sum_{i=1}^4 \frac{1}{F_i} \\ &= (7.2)^2 \left(\frac{1}{124.2} + \frac{1}{46.8} + \frac{1}{942.8} + \frac{1}{355.2} \right) \\ &= (51.84)(0.0333) = 1.73\end{aligned}\quad (8.10.1)$$

A table of reciprocals is useful in this calculation, since the four reciprocals can be added directly.

How many degrees of freedom has χ^2 ? Since all four deviations are the same except for sign, this suggests that χ^2 has only 1 *df.*, as was proved by Fisher. With 1 *df.*, table A 5 shows that a value of χ^2 greater than 1.73 occurs with probability about 0.20. The observed difference in proportion dead between the non-smokers and pipe smokers may well be due to sampling errors.

The above χ^2 has not been corrected for continuity. A correction is appropriate because the exact distribution of χ^2 in a 2×2 table is discrete. With the same four marginal totals, the two sets of results that are closest to our observed results are as follows:

(i)			(ii)		
118	53	171	116	55	171
949	349	1298	951	347	1298
1067	402		1067	402	
$f - F = \pm 6.2$			$f - F = \pm 8.2$		

Since the expected values do not change, the values $(f - F)$ are ± 6.2 in (i) and ± 8.2 in (ii), as against ± 7.2 in our data. Thus, in the exact distribution of χ^2 the values of $|f - F|$ jump by unity. The correction for continuity is made by deducting 0.5 from $|f - F|$. The formula for corrected χ^2 is

$$\begin{aligned}\chi_c^2 &= (|f - F| - 0.5)^2 \sum 1/F_i \\ &= (6.7)^2 (0.0333) = 1.49\end{aligned}\quad (8.10.2)$$

The corrected *P* is about 0.22, little changed in this example because the samples are large. In small samples the correction makes a substantial difference.

Some workers prefer an alternative formula for computing χ^2 . The 2×2 table may be represented in this way:

a	b	$a + b$
c	d	$c + d$
$a + c$	$b + d$	$N = a + b + c + d$

$$\chi_c^2 = \frac{N(|ad - bc| - N/2)^2}{(a + b)(c + d)(a + c)(b + d)} \quad (8.10.3)$$

The subtraction of $N/2$ represents the correction for continuity

In interpreting the results of these χ^2 tests in non-experimental studies, caution is necessary, particularly when χ^2 is significant. The two groups being compared may differ in numerous ways, some of which may be wholly or partly responsible for an observed significant difference. For instance, pipe smokers and non-smokers may differ to some extent in their economic levels, residence (urban or rural), and eating and drinking habits, and these variables may be related to the risk of dying. Before the investigator can claim that a significant difference is caused by the variable under study, it is his responsibility to produce evidence that disturbing variables of this type could not have produced the difference. Of course, the same responsibility rests with the investigator who has done a controlled experiment. But the device of randomization, and the greater flexibility which usually prevails in controlled experimentation, make it easier to ensure against misleading conclusions from disturbing influences.

EXAMPLE 8 10 1—In a study as to whether cancer of the breast tends to “run in families,” Murphy and Abbey (4) investigated the frequency of breast cancer found in relatives of (i) women with breast cancer, (ii) a comparison group of women without breast cancer. The data below, slightly altered for easy calculation, refer to the mothers of the subjects

		Breast Cancer in Subject		Total
		Yes	No	
Breast Cancer in Mother	Yes	7	3	10
	No	193	197	390
Total		200	200	400

Calculate χ^2 and P (i) without correction, (ii) with correction for continuity, for testing the null hypothesis that the frequency of cancer in mothers is the same in the two classes of subjects. Ans (i) $\chi^2 = 1.64$, $P = 0.20$ (ii) $\chi_c^2 = 0.92$, $P = 0.34$. Note that the correction for continuity always increases P , that is, makes the difference less significant.

EXAMPLE 8 10 2—In the previous example, verify that the alternative formula 8 10 3 for χ_c^2 gives the same result, by showing that χ_c^2 in 8 10 3 comes out as $12/13 = 0.92$.

EXAMPLE 8 10 3—Dr C H Richardson has furnished the following numbers of aphids (*Aphis rumicis* L) dead and alive after spraying with two concentrations of solutions of sodium oleate

	Concentration of Sodium Oleate (percentage)		Total
	0.65	1.10	
Dead	55	62	117
Alive	13	3	16
Total	68	65	133
Per Cent Dead	80.9	95.4	

Has the higher concentration given a significantly different per cent kill? Ans $\chi_c^2 = 5.31$, $P < 0.025$

EXAMPLE 8 10 4—In examining the effects of sprays in the control of codling moth injury to apples Hansberry and Richardson (5) counted the wormy apples on each of 48 trees Two trees sprayed with the same amount of lead arsenate yielded

A 2,130 apples 1 299 or 61% of which were injured

B 2,190 apples, 1,183 or 45% of which were injured

$\chi^2 = 21.16$ is conclusive evidence that the chance of injury was different in these two trees This result is characteristic of spray experiments For some unknown reasons, injuries under identical experimental treatments differ significantly Hence it is undesirable to compare sprays on single trees, because a difference in percentage of injured apples might be due to these unknown sources rather than to the treatments A statistical determination of the homogeneity or heterogeneity of experimental material under identical conditions, sometimes called a *test of technique*, is often worthwhile, particularly in new fields of research.

EXAMPLE 8 10 5—Prove that formulas 8 10 2 and 8 10 3 for χ_c^2 are the same, by showing that

$$|f - F| = |ad - bc|/N$$

$$\Sigma(1/F) = N^2/(a + b)(c + d)(a + c)(b + d)$$

8.11—Test of the independence of two attributes. The preceding test is sometimes described as a test of the independence of two attributes A sample of people of a particular ethnic type might be classified into two classes according to hair color and also into two classes according to color of eyes We might ask “are color of hair and color of eyes independent?” Similarly, the numerical example in the previous section might be referred to as a test of the question “Is the risk of dying independent of smoking habit?”

In this way of speaking, the word “independent” carries the same meaning as it does in Rule 3 in the theory of probability Let p_A be the probability that a member of a population possesses attribute A , and p_B the probability that he possesses attribute B If the attributes are independent, the probability that he possesses both attributes is $p_A p_B$ Thus, on the null hypothesis of independence, the probabilities in the four cells of the 2×2 contingency table are as follows

		Attribute A		Total
		(1) Present	(2) Absent	
Attribute B	(1) Present	$p_A p_B$	$q_A p_B$	p_B
	(2) Absent	$p_A q_B$	$q_A q_B$	q_B
Total		p_A	q_A	1

Two points emerge from this table The null hypothesis can be tested either by comparing the proportions of cases in which B is present in columns (1) and (2), or by comparing the proportions of cases in which A is present in rows (1) and (2) These two χ^2 tests are exactly the same. This is not obvious from the original expressions (8 10 1) and (8 10 2) given for χ^2 and χ_c^2 , but expression (8 10 3) makes it clear that the statement holds

Secondly, the table provides a check on the rule given for calculating the expected number in any cell. In a single sample of size N , we expect to find $Np_A p_B$ members possessing both A and B . The sample total in column (1) will be our best estimate of Np_A , while that in row (1) similarly estimates Np_B . Thus the rule, (column total)(row total)/(grand total) gives $(N\hat{p}_A)(N\hat{p}_B)/N = N\hat{p}_A\hat{p}_B$ as required.

8.12—A test by means of the normal deviate z . The null hypothesis can also be tested by computing a normal deviate z , derived from the normal approximation to the binomial. The z and χ^2 tests are identical. Many investigators prefer the z form, because they are primarily interested in the size of the difference $\hat{p}_1 - \hat{p}_2$ between the proportions found in two independent samples. For illustration, we repeat the data from table 8.10.1.

TABLE 8.12.1
MEN CLASSIFIED BY SMOKING HABIT AND MORTALITY IN SIX YEARS

	Sample (1) Non-smokers	Sample (2) Pipe Smokers	Total
Dead	117	54	171
Alive	950	348	1,298
Total	$n_1 = 1,067$	$n_2 = 402$	1,469
Proportion dead	$\hat{p}_1 = 0.1097$	$\hat{p}_2 = 0.1343$	$\hat{p} = 0.1164$

Since $\hat{p}_1 = 0.1097$ and $\hat{p}_2 = 0.1343$ are approximately normally distributed, their difference $\hat{p}_1 - \hat{p}_2$ is also approximately normally distributed. The variance of this difference is the sum of the two variances (section 4.7).

$$V(\hat{p}_1 - \hat{p}_2) = \sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

Under the null hypothesis, $p_1 = p_2 = p$, so that $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed with mean 0 and standard error

$$\sqrt{\left\{ \frac{pq}{n_1} + \frac{pq}{n_2} \right\}}$$

The null hypothesis does not specify the value of p . As an estimate, we naturally use $\hat{p} = 0.1164$ as given by the combined samples. Hence, the normal deviate z is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.1097 - 0.1343}{\sqrt{\left\{(0.1164)(0.8836)\left(\frac{1}{1067} + \frac{1}{402}\right)\right\}}} = \frac{-0.0246}{0.01877} = -1.31$$

In the normal table, ignoring the sign of z , we find $P = 0.19$, in agreement with the value found by the original χ^2 test.

To correct z for continuity, subtract $\frac{1}{2}$ from the numerator of the *larger* proportion (in this case \hat{p}_2) and add $\frac{1}{2}$ to the numerator of the *smaller* proportion. Thus, instead of \hat{p}_2 we use $\hat{p}_2' = 53.5/402 = 0.1331$ and instead of \hat{p}_1 we use $\hat{p}_1' = 117.5/1067 = 0.1101$. The denominator of z_c remains the same, giving $z_c = (0.1101 - 0.1331)/0.01877 = -1.225$. You may verify that, apart from rounding errors, $z^2 = \chi^2$ and $z_c^2 = \chi_c^2$.

If the null hypothesis has been rejected and you wish to find confidence limits for the population difference $p_1 - p_2$, the standard error of $\hat{p}_1 - \hat{p}_2$ should be computed as

$$\sqrt{\left\{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}\right\}}$$

The *s.e.* given by the null hypothesis is no longer valid. Often the change is small, but it can be material if n_1 and n_2 are very unequal.

EXAMPLE 8.12.1—Apply the z test and the z_c test to the data on breast cancer given in example 8.10.1 and verify that $z^2 = \chi^2$ and $z_c^2 = \chi_c^2$. Note. when calculating z or z_c it is often more convenient to express \hat{p}_1 , \hat{p}_2 and \hat{p} as percentages. Just remember that in this event, $\hat{q} = 100 - \hat{p}$.

EXAMPLE 8.12.2—In 1943 a sample of about 1 in 1,000 families in Iowa was asked about the canning of fruits or vegetables during the preceding season. Of the 392 rural families, 378 had done canning, while of the 300 urban families, 274 had canned. Calculate 95% confidence limits for the difference in the percentages of rural and urban families who had canned. Ans. 1.42% and 8.78%.

The preceding χ^2 and z methods are approximate, the approximation becoming poorer as the sample size decreases. Fisher (14) has shown how to compute an exact test of significance. For accurate work the exact test should be used if (i) the total sample size N is less than 20, or (ii) if N lies between 20 and 40 and the smallest expected number is less than 5. For those who encounter these conditions frequently, reference (15), which gives tables of the exact tests covering these cases, is recommended.

8.13—Sample size for comparing two proportions. The question: How large a sample do I need? is naturally of great interest to investigators. For comparing two means, an approach that is often helpful was given in section 4.13, p. 111. This should be reviewed carefully, since the same principle applies to the comparison of two proportions. The approach assumes that it is planned to make a test of significance of the difference

between the two proportions, and that future actions will depend on whether the test shows a significant difference or not. Consequently, if the true difference $p_2 - p_1$ is as large as some amount δ chosen by the investigator, he would like the test to have a high probability P' of declaring a significant result.

For two independent samples, formula (4.13.1) (p. 113) for n , the size of each sample, can be applied. Put $\delta = p_2 - p_1$ and $\sigma_D^2 = (p_1q_1 + p_2q_2)$. This gives

$$n = (Z_\alpha + Z_\beta)^2(p_1q_1 + p_2q_2)/(\delta)^2 \quad (8.13.1)$$

where Z_α is the normal deviate corresponding to the significance level to be used in the test, $\beta = 2(1 - P')$, and Z_β is the normal deviate corresponding to the *two-tailed* probability β . Table 4.13.1 gives $(Z_\alpha + Z_\beta)^2$ for the commonest values of α and β . In using this formula, we substitute the best advance estimate of $(p_1q_1 + p_2q_2)$ in the numerator.

For instance, suppose that a standard antibiotic has been found to protect about 50% of experimental animals against a certain disease. Some new antibiotics become available that seem likely to be superior. In comparing a new antibiotic with the standard, we would like a probability $P' = 0.9$ of finding a significant difference in a one-tailed test at the 5% level if the new antibiotic will protect 80% of the animals in the population. For these conditions, table 4.13.1 gives $(Z_\alpha + Z_\beta)^2$ as 8.6. Hence

$$n = (8.6)\{(50)(50) + (80)(20)\}/(30)^2 = 39.2$$

Thus, 40 animals should be used for each antibiotic.

Some calculations of this type will soon convince you of the sad fact that large samples are necessary to detect small differences between two percentages. When resources are limited, it is sometimes wise, before going ahead with the experiment, to calculate the probability that a significant result will be found. Suppose that an experimenter is interested in the values $p_1 = 0.8$, $p_2 = 0.9$, but cannot make $n > 100$. If formula (8.13.1) is solved for Z_β , we find

$$Z_\beta = \frac{(p_2 - p_1)\sqrt{n}}{\sqrt{p_1q_1 + p_2q_2}} - Z_\alpha = \frac{(0.1)(10)}{0.5} - Z_\alpha = 2 - Z_\alpha$$

If he intends a two-tailed 5% test $Z_\alpha \doteq 2$, so that $Z_\beta \doteq 0$. This gives $\beta = 1$ and $P' = 1 - \beta/2 = 0.5$. The proposed experiment has only a 50-50 chance of finding a significant difference in this situation.

Formula (8.13.1), although a large-sample approximation, should be accurate enough for practical use, since there is usually some uncertainty about the values of \hat{p}_1 and \hat{p}_2 to insert in the formula. Reference (6) gives tables of n based on a more accurate approximation.

EXAMPLE 8.13.1—One difficulty in estimating sample size in biological work is that the proportions given by a standard treatment may vary over time. An experimenter has found that his standard treatment has a failure rate lying between $p_1 = 30\%$ and $p_1 = 40\%$. With a new treatment whose failure rate is 20% lower than the standard, what sample sizes are needed to make $P' = 0.9$ in a two-tailed 5% test? Ans. $n = 79$ when $p_1 = 30\%$ and $n = 105$ when $p_1 = 40\%$.

EXAMPLE 8.13.2—In planning the 1954 trial of the Salk poliomyelitis vaccine (7), the question of sample size was critical, since it was unlikely that the trial could be repeated and since an extremely large sample of children would obviously be necessary. Various estimates of sample size were therefore made. In one of these it was assumed that the probability that an unprotected child would contract paralytic polio was 0.0003 , or 0.03% . If the vaccine was 50% effective (that is, decreased this probability to 0.00015 , or 0.015%), it was desired to have a 90% chance of finding a 5% significance difference in a two-tailed test. How many children are required? Ans. $210,000$ in each group (vaccinated and unprotected)

EXAMPLE 8.13.3—An investigator has $p_1 = 0.4$, and usually conducts experiments with $n = 25$. In a one-tailed test at the 5% level, what is the chance of obtaining a significant result if (i) $p_2 = 0.5$, (ii) $p_2 = 0.6$? Ans. (i) 0.18 , (ii) 0.42 .

8.14—The Poisson distribution. As we have seen, the binomial distribution tends to the normal distribution as n increases for any fixed value of p . The value of n needed to make the normal approximation a good one depends on the values of p , this value being smallest when $p = 1/2$. For $p < 1/2$, a general rule, usually conservative, is that the normal approximation is adequate if the mean $\mu = np$ is greater than 15 .

In many applications, however, we are studying rare events, so that even if n is large, the mean np is much less than 15 . The binomial distribution then remains noticeably skew and the normal approximation is unsatisfactory. A different approximation for such cases was developed by S. D. Poisson (8). He worked out the limiting form of the binomial distribution when n tends to infinity and p tends to zero at the same time, in such a way that $\mu = np$ is constant. The binomial expression for the probability of r successes tends to the simpler form,

$$P(r) = \frac{\mu^r}{r!} e^{-\mu} \quad r = 0, 1, 2, \dots,$$

where $e = 2.71828$ is the base of natural logarithms. The initial terms in the Poisson distribution are:

$$P(0) = e^{-\mu} : P(1) = \mu e^{-\mu} : P(2) = \frac{\mu^2}{2} e^{-\mu} : P(3) = \frac{\mu^3}{(2)(3)} e^{-\mu}$$

Table 8.14.1 shows in column (1) the Poisson distribution for $\mu = 1$. The distribution is markedly skew. The mode (highest frequency) is at either 0 or 1, these two having the same probability when $\mu = 1$. To give an idea of the way in which the binomial tends to approach the Poisson, column (2) shows the binomial distribution for $n = 100$, $p = 0.01$, and column (3) the binomial for $n = 25$, $p = 0.04$, both of these having $np = 1$. The agreement with the Poisson is very close for $n = 100$ and

TABLE 8.14.1
THE POISSON DISTRIBUTION FOR $\mu = 1$ COMPARED WITH THE BINOMIAL
DISTRIBUTIONS FOR $n = 100, p = 0.01$ AND $n = 25, p = 0.04$

r	Relative Frequencies		
	(1) Poisson 1	(2) Binomial $n = 100, p = 0.01$	(3) Binomial $n = 25, p = 0.04$
0	0.3679	0.3660	0.3604
1	0.3679	0.3697	0.3754
2	0.1839	0.1849	0.1877
3	0.0613	0.0610	0.0600
4	0.0153	0.0149	0.0137
5	0.0031	0.0029	0.0024
6	0.0005	0.0005	0.0003
≥ 7	0.0001	0.0001	0.0000
Total	1.0000	1.0000	0.9999

quite close for $n = 25$. Tables of individual and cumulative terms of the Poisson are given in (9) and of individual terms up to $\mu = 15$ in (10).

The fitting of a Poisson distribution to a sample will be illustrated by the data (11) in table 8.14.2. These show the number of noxious weed seeds in 98 sub-samples of *Phleum praetense* (meadow grass). Each sub-sample weighed 1/4 ounce, and of course contained many seeds, of which only a small percentage were noxious. The first step is to compute the sample mean.

$$\hat{\mu} = (\Sigma fr) / (\Sigma f) = 296/98 = 3.0204 \text{ noxious seeds per sub-sample}$$

TABLE 8.14.2
DISTRIBUTION OF NUMBER OF NOXIOUS WEED SEEDS FOUND IN $N = 98$
SUB-SAMPLES, WITH FITTED POISSON DISTRIBUTION

Number of Noxious Seeds r	Frequency f	Poisson Multipliers	Expected Frequency
0	3	$1 \approx 1.0000$	4.781
1	17	$\hat{\mu} = 3.0204$	14.440
2	26	$\hat{\mu}/2 = 1.5102$	21.807
3	16	$\hat{\mu}/3 = 1.0068$	21.955
4	18	$\hat{\mu}/4 = 0.7551$	16.578
5	9	$\hat{\mu}/5 = 0.6041$	10.015
6	3	$\hat{\mu}/6 = 0.5034$	5.042
7	5	$\hat{\mu}/7 = 0.4315$	2.176
8	0	$\hat{\mu}/8 = 0.3756$	0.817
9	1	$\hat{\mu}/9 = 0.3356$	0.274
10	0	$\hat{\mu}/10 = 0.3020$	0.083
11 or more	0	$\hat{\mu}/11 = 0.2746$	0.030
Total	98		97.998

Next, calculate the successive terms of the Poisson distribution with mean $\hat{\mu}$. The expected number of sub-samples with 0 seeds is $Ne^{-\hat{\mu}} = (98)(e^{-3.0204})$. A table of natural logs gives $e^{-3.0204} = 1/20.5$, and $98/20.5 = 4.781$. Next, form a column of the successive multipliers $1, \hat{\mu}, \hat{\mu}/2, \dots$ as shown in table 8.14.2, recording each to at least four significant digits. The expected number of sub-samples with $r = 1$ is $(4.781)(\hat{\mu}) = 14.440$. Similarly, the expected number with $r = 2$ is $(14.440)(\hat{\mu}/2) = (14.440)(1.5102) = 21.807$, and so on. The agreement between observed and expected frequencies seems good except perhaps for $r = 2$ and $r = 3$, which have almost equal expected numbers but have observed numbers 26 and 16. A test of the discrepancies between observed and expected numbers (section 9.6), shows that these can well be accounted for by sampling errors.

Two important properties hold for a Poisson variate. The variance of the distribution is equal to its mean, μ . This would be expected, since the binomial variance, npq , tends to np when q tends to 1. Secondly, if a series of *independent* variates X_1, X_2, X_3, \dots each follow Poisson distributions with means $\mu_1, \mu_2, \mu_3, \dots$, their sum follows a Poisson distribution with mean $(\mu_1 + \mu_2 + \mu_3 + \dots)$.

In the inspection and quality control of manufactured goods, the proportion of defective articles in a large lot should be small. Consequently, the number of defectives in the lot might be expected to follow a Poisson distribution. For this reason, the Poisson distribution plays an important role in the development of plans for inspection and quality control. Further, the Poisson is often found to serve remarkably well as an approximation when μ is small, even if the value of n is ill-defined and if both n and p presumably vary from one sample to another. A much-quoted example of a good fit of a Poisson distribution, due to Bortkewitch, is to the number of men in a Prussian army corps who were killed during a year by the kick of a horse. He had $N = 200$ observations, one for each of 10 corps for each of 20 years. On any given day, some men were exposed to a small probability of being kicked, but is not clear what value n has, nor that p would be constant.

The Poisson distribution can also be developed by reasoning quite unrelated to the binomial. Suppose that signals are being transmitted, and that the probability that a signal reaches a given point in a tiny time-interval τ is $\lambda\tau$, irrespective of whether previous signals have arrived recently or not. Then the number of signals arriving in a finite time-interval of length T may be shown to follow a Poisson distribution with mean λT (example 8.14.4). Similarly, if particles are distributed at random in a liquid with density λ per unit volume, the number found in a sample of volume V is a Poisson variable with mean λV . From these illustrations it is not surprising that the Poisson distribution has found applications in many fields, including communications theory and the estimation of bacterial densities.

EXAMPLE 8.14.1 — $n = 1,000$ independent trials are made of an event with probability

301 at each trial. Give approximate results for the chances that (i) the event does not happen, (ii) the event happens twice, (iii) the event happens at least five times. Ans (i) 0.368, (ii) 0.184, (iii) 0.0037

EXAMPLE 8 14 2—A. G. Arbous and J. E. Kerrich (12) report the numbers of accidents sustained during their first year by 155 engine shunters aged 31–35, as follows

no. of accidents	0	1	2	3	4 or more
no. of men	80	61	13	1	0

Fit a Poisson distribution to these data. Note the data were obtained as part of a study of accident proneness. If some men are particularly liable to accidents, this would imply that the Poisson would not be a good fit, since p would vary from man to man.

EXAMPLE 8 14 3—Student (13) counted the number of yeast cells on each of 400 squares of a hemacytometer. In two independent samples, each of which gave a satisfactory fit to a Poisson distribution, the total numbers of cells were 529 and 720. (i) Test whether these totals are estimates of the same quantity, or in other words whether the density of yeast cells per square is the same in the two populations. (ii) Compute 95% limits for the difference in density per square. Ans (i) $z = 5.41$. P very small. (ii) 0.30 to 0.65. Note the normal approximation to the Poisson distribution, or to the difference between two independent Poisson variates, may be used when the observed numbers exceed 15.

EXAMPLE 8 14 4—The Poisson process formula for the number of signals arriving in a finite time-interval T requires one result in calculus, but is otherwise a simple application of probability rules. Let $P(r, T + \tau)$ denote the probability that exactly r signals have arrived in the interval from time 0 to the end of time $(T + \tau)$. This event can happen in one of two mutually exclusive ways: (i) $(r - 1)$ signals have arrived by time T , and one arrives in the small interval τ . The probability of these two events is $\lambda\tau P(r - 1, T)$. (ii) r signals have already arrived by time T , and none arrives in the subsequent interval τ . The probability of these two events is $(1 - \lambda\tau)P(r, T)$. The interval τ is assumed so small that more than one signal cannot arrive in this interval. Hence,

$$P(r, T + \tau) = \lambda\tau P(r - 1, T) + (1 - \lambda\tau)P(r, T)$$

Rearranging, we have

$$\{P(r, T + \tau) - P(r, T)\}/\tau = \lambda\{P(r - 1, T) - P(r, T)\}$$

Letting τ tend to zero, we get $\partial P(r, T)/\partial T = \lambda\{P(r - 1, T) - P(r, T)\}$. By differentiating, it will be found that $P(r, T) = e^{-\lambda T}(\lambda T)^r/r!$ satisfies this equation.

REFERENCES

1. F. MOSTELLER, R. E. K. ROURKE, and G. B. THOMAS, JR. *Probability With Statistical Applications*. Addison-Wesley, Reading, Mass (1961).
2. Data made available by Dr. Martin Frobisher.
3. E. W. R. BEST, C. B. WALKER, and P. M. BAKER, *et al.* A Canadian Study on Smoking and Health (Final Report). Dept. of National Health and Welfare, Canada (1966).
4. D. P. MURPHY and H. ABBEY. *Cancer in Families*. Harvard University Press, Cambridge (1959).
5. T. R. HANSBERRY and C. H. RICHARDSON. *Iowa State Coll. J. Sci.*, 10, 27 (1935).
6. W. G. COCHRAN and G. M. COX. *Experimental Designs*. Wiley, New York, 2nd ed., p. 17 (1957).
7. T. J. FRANCIS, *et al.* *Evaluation of the 1954 Field Trial of Poliomyelitis Vaccine*. Edwards Bros., Inc., Ann Arbor (1957).
8. S. D. POISSON. *Recherches sur la probabilité des jugements*. Paris (1837).

- 9 E C MOLINA *Poisson's Exponential Binomial Limit* Van Nostrand, New York (1942)
- 10 E S PEARSON and H O HARTLEY *Biometrika Tables for Statisticians*, Vol I Cambridge University Press, Cambridge, England, 2nd ed (1966)
- 11 C W LEGGATT *Comptes rendus de l'association internationale d'essais de semences*, 5 27 (1935)
- 12 A G ARBOUS and J E KERRICH *Biometrics*, 7 340 (1951)
- 13 "Student " *Biometrika*, 5 351 (1907)
- 14 R A FISHER *Statistical Methods for Research Workers* §21.02 Oliver and Boyd, Edinburgh
- 15 D J FINNEY, R LATSCHA, B M BENNETT, and P Hsu *Tables for Testing Significance in a 2×2 Contingency Table* Cambridge University Press, New York (1963)
- 16 National Bureau of Standards *Tables of the Binomial Probability Distribution* App Math Series 6 (1950)
- 17 Annals of the Computation Laboratory *Tables of the Cumulative Binomial Probability Distribution* Harvard University, Vol 35 (1955)

Attribute data with more than one degree of freedom

9.1—Introduction. In chapter 8 the discussion of attribute data was confined to the cases in which the population contains only two classes of individuals and in which only one or two populations have been sampled. We now extend the discussion to populations classified into more than two classes, and to samples drawn from more than two populations. Section 9.2 considers the simplest situation in which the expected numbers in the classes are completely specified by the null hypothesis.

9.2—Single classifications with more than two classes. In crosses between two types of maize, Lindstrom (1) found four distinct types of plants in the second generation. In a sample of 1,301 plants, there were

$$\begin{array}{rcl} f_1 & = & 773 \text{ green} \\ f_2 & = & 231 \text{ golden} \\ f_3 & = & 238 \text{ green-striped} \\ f_4 & = & 59 \text{ golden-green-striped} \\ & & \hline & & 1301 \end{array}$$

According to a simple type of Mendelian inheritance, the probabilities of obtaining these four types of plants are $9/16$, $3/16$, $3/16$, and $1/16$, respectively. We select this as the null hypothesis.

The χ^2 test in chapter 8 is applicable to any number of classes. Accordingly, we calculate the numbers of plants that would be expected in the four classes if the null hypothesis were true. These numbers, and the deviations ($f - F$), are shown below.

$$\begin{array}{rclcl} F_1 = (9/16)(1301) & = & 731.9 & : & f_1 - F_1 & +41.1 \\ F_2 = (3/16)(1301) & = & 243.9 & : & f_2 - F_2 & = -12.9 \\ F_3 = (3/16)(1301) & = & 243.9 & : & f_3 - F_3 & = -5.9 \\ F_4 = (1/16)(1301) & = & 81.3 & : & f_4 - F_4 & = -22.3 \\ & & \hline & & 1301.0 & & & 0.0 \end{array}$$

Substituting in the formula for chi-square,

$$\begin{aligned}\chi^2 &= \Sigma(f - F)^2/F \\ \chi^2 &= \frac{(41.1)^2}{731.9} + \frac{(-12.9)^2}{243.9} + \frac{(-5.9)^2}{243.9} + \frac{(-22.3)^2}{81.3} \\ &= 2.31 + 0.68 + 0.14 + 6.12 \\ &= 9.25\end{aligned}$$

In a test of this type, the number of degrees of freedom in χ^2 = (Number of classes) - 1 = 4 - 1 = 3. To remember this rule, note that there are four deviations, one for each class. However, the sum of the four deviations, 41.1 - 12.9 - 5.9 - 22.3, is zero. Only three of the deviations can vary at will, the fourth being fixed as zero minus the sum of the first three.

Is χ^2 as large as 9.25, with $d.f. = 3$, a common event in sampling from the population specified by the null hypothesis 9 : 3 : 3 : 1, or is it a rare one? For the answer, refer to the χ^2 table (table A 5, p. 550), in the line for 3 $d.f.$ You will find that 9.25 is beyond the 5% point, near the 2.5% point. On this evidence the null hypothesis would be rejected.

When there are more than two classes, this χ^2 test is usually only a first step in the examination of the data. From the test we have learned that the deviations between observed and expected numbers are too large to be reasonably attributed to sampling fluctuations. But the χ^2 test does not tell us in what way the observed and expected numbers differ. For this, we must look at the individual deviations and their contributions to χ^2 . Note that the first class, (green), gives a large positive deviation +41.1 and is the only class giving a positive deviation. Among the other classes, the last class (golden-green-striped) gives the largest deviation, -22.3, and the largest contribution to χ^2 , 6.12 out of a total of 9.25. Lindstrom commented that the deviations could be largely explained by a physiological cause, namely the weakened condition of the last three classes due to their chlorophyll abnormality. He pointed out in particular that the last class (golden-green-striped) was not very vigorous.

To illustrate the type of subsequent analysis that is often necessary with more than two classes, let us examine whether the data are consistent with the weaker hypothesis that the numbers in the first three classes are in the predicted Mendelian ratios 9 : 3 : 3. If so, one interpretation of the results is that the significant value of χ^2 can be attributed to poor survivorship of the golden-green-striped class.

The 9 : 3 : 3 hypothesis is tested by a χ^2 test applied to the first three classes. The calculations appear in table 9.2.1.

In the first class, $F_1 = (0.6)(1242) = 745.2$, and so on. The value of χ^2 is now 2.70, with 3 - 1 = 2 $d.f.$ Table A 5 shows that the probability is about 0.25 of obtaining a χ^2 as large as this when there are 2 $d.f.$

We can also test whether the last class (golden-green-striped) has a frequency of occurrence significantly less than would be expected from its Mendelian probability 1/16. For this we observe that 1242 plants fell

TABLE 9.2.1
TEST OF THE MENDELIAN HYPOTHESIS IN THE FIRST THREE CLASSES

Class	f	Hypothetical Probability	F	$f - F$	$(f - F)^2/F$
green	773	$9/15 = 0.6$	745.2	+27.8	1.04
golden	231	$3/15 = 0.2$	248.4	-17.4	1.22
green-striped	238	$3/15 = 0.2$	248.4	-10.4	0.44
Total	1242	$15/15 = 1$	1242.0	0.0	2.70

into the first three classes, which have total probability 15/16, as against 59 plants in the fourth class, with probability 1/16. The corresponding expected numbers are 1219.7 and 81.3. In this case the χ^2 test reduces to that given in section 8.8 for testing a theoretical binomial proportion. We have

$$\begin{aligned}\chi^2 &= \frac{(1242 - 1219.7)^2}{1219.7} + \frac{(59 - 81.3)^2}{81.3} \\ &= \frac{(+22.3)^2}{1219.7} + \frac{(-22.3)^2}{81.3} = 6.53,\end{aligned}$$

with 1 *df.* The significance probability is close to the 1% level.

To summarize, the high value of χ^2 obtained initially, 9.25 with 3 *df.*, can be ascribed to a deficiency in the number of golden-green-striped plants, the other three classes not deviating abnormally from the Mendelian probabilities. (There may be also, as Lindstrom suggests, some deficiencies in the second and third classes relative to the first class, which would show up more definitely in a larger sample.)

This device of making comparisons among sub-groups of the classes is useful in two situations. Sometimes, especially in exploratory work, the investigator has no clear ideas about the way in which the numbers in the classes will deviate from the initial null hypothesis: indeed, he may consider it likely that his first χ^2 test will support the null hypothesis. The finding of a significant χ^2 should be followed, as in the above example, by inspection of the deviations to see what can be learned from them. This process may lead to the construction of new hypotheses that are tested by further χ^2 tests among sub-groups of the classes. Conclusions drawn from this analysis must be regarded as tentative, because the new hypotheses were constructed after seeing the data and should be strictly tested by gathering new data.

In the second situation the investigator has some ideas about the types of departure that the data are likely to show from the initial null hypothesis; in other words, about the nature of the alternative hypothesis. The best procedure is then to construct tests aimed specifically at these types of departure. Often, the initial χ^2 test is omitted in this situation. This approach will be illustrated in later sections.

When calculating χ^2 with more than 1 *df.*, it is not worthwhile to

make a correction for continuity. The exact distribution of χ^2 is still discrete, but the number of different possible values of χ^2 is usually large, so that the correction, when properly made, produces only a small change in the significance probability.

EXAMPLE 9.2.1—In 193 pairs of Swedish twins (2), 56 were of type *MM* (both male), 72 of the type *MF* (one male, one female), and 65 of the type *FF*. On the hypothesis that a twin is equally likely to be a boy or a girl and that the sexes of the two members of a twin pair are determined independently, the probabilities of *MM*, *MF*, and *FF* pairs are $1/4$, $1/2$, $1/4$, respectively. Compute the value of χ^2 and the significance probability. Ans. $\chi^2 = 13.27$, with 2 *df*. $P < 0.005$.

EXAMPLE 9.2.2—In the preceding example we would expect the null hypothesis to be false for two reasons. The probability that a twin is male is not exactly $1/2$. This discrepancy produces only minor effects in a sample of size 193. Secondly, identical twins are always of the same sex. The presence of identical twins decreases the probability of *MF* pairs and increases the probabilities of *MM* and *FF* pairs. Construct χ^2 tests to answer the questions: (i) Are the relative numbers of *MM* and *FF* pairs (ignoring the *MF* pairs) in agreement with the null hypothesis? (ii) Are the relative numbers of twins of like sex (*MM* and *FF* combined) and unlike sex (*MF*) in agreement with the null hypothesis? Ans. (i) χ^2 (uncorrected) = 0.67, with 1 *df*. $P > 0.25$, (ii) $\chi^2 = 12.44$, with 1 *df*. P very small. The failure of the null hypothesis is due, as anticipated, to an excess of twins of like sex.

EXAMPLE 9.2.3—In section 1.14, 230 samples from binomial distributions with known p were drawn, and χ^2 was computed from each sample. The observed and expected numbers of χ^2 values in each of seven classes (taken from table 1.14.1) are as follows:

Obs.	57	59	62	32	14	3	3	230
Exp.	57.5	57.5	57.5	34.5	11.5	9.2	2.3	230.0

Test whether the deviations of observed from expected numbers are of a size that occurs frequently by chance. Ans. $\chi^2 = 5.50$, *df* = 6. P about 0.5.

EXAMPLE 9.2.4—In the Lindstrom example in the text, we had χ_3^2 (3 *df*.) = 9.25. This was followed by χ_2^2 (2 *df*.) = 2.70, which compared the first three classes, and $\chi_1^2 = 6.53$, which compared the combined first three classes with the fourth class. Note that $\chi_2^2 + \chi_1^2 = 9.23$, while $\chi_3^2 = 9.25$. In examples 9.2.1 and 9.2.2, $\chi_1^2 = 13.27$, while the sum of the two 1-*df*. chi-squares is $0.67 + 12.44 = 13.11$. When a classification is divided into sub-groups and a χ^2 is computed within each sub-group, plus a χ^2 which compares the total frequencies in the sub-groups, the *df*. add up to the *df*. in the initial χ^2 , but the values of χ^2 do not add up exactly to the initial χ^2 . They usually add to a value that is fairly close, and worth noting as a clue to mistakes in calculation.

9.3—Single classifications with equal expectations. Often, the null hypothesis specifies that all the classes have equal probabilities. In this case, χ^2 has a particularly simple form. As before, let f_i denote the observed frequency in the *i*th class, and let $n = \sum f_i$ be the total size of sample. If there are *k* classes, the null hypothesis probability that a member of the population falls into any class is $p = 1/k$. Consequently, the expected frequency F_i in any class is $np = n/k = \bar{f}$, the mean of the f_i . Thus,

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - F_i)^2}{F_i} = \sum_{i=1}^k \frac{(f_i - \bar{f})^2}{\bar{f}},$$

with $(k - 1)$ *df*.

This test is applied to any new table of random numbers. The basic property of such a table is that each digit has a probability $1/10$ of being chosen at each draw. To illustrate the test, the frequencies of the first 250 digits in the random number table A 1 are as follows:

Digit	0	1	2	3	4	5	6	7	8	9	Total
f_i	22	24	28	23	18	33	29	17	31	25	250

Only 17 *sevens* and 18 *fours* have appeared, as against 31 *eights* and 33 *fives*. The mean frequency $\bar{f} = 25$. Thus, by the usual shortcut method of computing the sum of squares of deviations, $\Sigma(f_i - \bar{f})^2$, given in section 2.10,

$$\chi^2 = \frac{1}{25} [(22)^2 + (24)^2 + \dots + (25)^2 - (250)^2/10] = 10.08,$$

with 9 *d.f.* Table A 5 shows that the probability of a χ^2 as large as this lies between 0.5 and 0.3: χ^2 is not unusually large.

This test can be related to the Poisson distribution. Suppose that the f_i are the numbers of occurrences of some rare event in a series of k independent samples. The null hypothesis is that the f_i all follow Poisson distributions with the same mean μ . Then, as shown by Fisher, the quantity $\Sigma(f_i - \bar{f})^2/\bar{f}$ is distributed approximately as χ^2 with $(k - 1)$ *d.f.* To go a step further, the test can be interpreted as a comparison of the observed variance of the f_i with the variance that would be expected from the Poisson distribution. In the Poisson distribution, the variance equals the mean μ , of which the sample estimate is \bar{f} . The observed variance among the f_i is $s^2 = \Sigma(f_i - \bar{f})^2/(k - 1)$. Hence

$$\chi^2 = (k - 1) (\text{observed variance})/(\text{Poisson variance})$$

This χ^2 test is sensitive in detecting the alternative hypothesis that the f_i follow independent Poisson distributions with *different* means μ_i . Under this alternative, the expected value of χ^2 may be shown to be, approximately,

$$E(\chi^2) \doteq (k - 1) + \sum_{i=1}^k (\mu_i - \bar{\mu})^2/\bar{\mu},$$

where $\bar{\mu}$ is the mean of the μ_i . If the null hypothesis holds, $\mu_i = \bar{\mu}$ and χ^2 has its usual average value $(k - 1)$. But any differences among the μ_i increase the expected value of χ^2 and tend to make it large. The test is sometimes called a variance test of the homogeneity of the Poisson distribution.

Sometimes the number of Poisson samples k is large. When computing the variance, time may be saved by grouping the observations, particularly if they take only a limited number of distinct values. To avoid confusion in our notation, denote the numbers of occurrences by y_i in-

stead of f_i , since we have used f 's in previous chapters to denote the frequencies found in a grouped sample. In this notation,

$$\chi^2 = \sum_{i=1}^k \frac{(y_i - \bar{y})^2}{\bar{y}} = \sum_{j=1}^m \frac{f_j(y_j - \bar{y})^2}{\bar{y}} = \frac{1}{\bar{y}} \left[\sum_{j=1}^m f_j y_j^2 - (\sum f_j y_j)^2 / \sum f_j \right],$$

where the second sum is over the m distinct values of y , and f_j is the frequency with which the j th value of y appears in the sample. The $d.f.$ are, as before, $(k - 1)$.

If the $d.f.$ in χ^2 lie beyond the range covered in table A 5, calculate the approximate normal deviate

$$Z = \sqrt{2\chi^2} - \sqrt{2(d.f.) - 1} \quad (9.3.1)$$

The significance probability is read from the normal table, using only one tail. For an illustration of this case, see examples 9.3.2 and 9.3.3.

EXAMPLE 9 3 1—In 1951, the number of babies born with a harelip in Birmingham, England, are quoted by Edwards (3) as follows

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
Number	8	19	11	12	16	8	7	5	8	3	8	8

Test the null hypothesis that the probability of a baby with harelip is the same in each month. Ans. $\chi^2 = 23.5$, $d.f. = 11$. P between 0.025 and 0.01. Strictly, the variable that should be examined in studies of this type is the ratio (number of babies with harelip)/(total number of babies born), because even if this ratio is constant from month to month, the actual number of babies with harelip will vary if the total number born varies. Edwards points out that in these data the total number varies little and shows no relation to the variation in number with harelip. He proceeds to fit the above data by a periodic (cosine) curve, which indicates a maximum in March.

EXAMPLE 9 3 2—Leggatt (4) counted the number of seeds of the weed *potentilla* found in 98 quarter-ounce batches of the grass *Phleum pratense*. The 98 numbers varied from 0 to 7, and were grouped into the following frequency distribution

Number of seeds	y_j	0	1	2	3	4	5	6	7	Total
Number of batches	f_j	37	32	16	9	2	0	1	1	98

Calculate $\chi^2 = \sum f_j(y_j - \bar{y})^2 / \bar{y}$. Ans. $\chi^2 = 145.4$, with 97 $d.f.$ From table A 5, with 100 $d.f.$, P is clearly less than 0.005. The high value of χ^2 is due to the batches with six and seven seeds.

EXAMPLE 9 3 3—Compute the significance probability in the preceding example by finding the normal deviate Z given by equation 9.3.1. Ans. $z = 3.16$, $P = 0.0008$. The correct probability, found from a larger table of χ^2 , is $P = 0.0010$.

9.4—Additional tests. As in section 9.2, the χ^2 test for the Poisson distribution can be supplemented or replaced by other tests directed more specifically against the type of alternative hypothesis that the investigator has in mind. If it is desired to examine whether a rare meteorological event occurs more frequently in the summer months, we might compare

the total frequency in June, July, and August with the total frequency in the rest of the year, the null hypothesis probabilities being very close to $1/4$ and $3/4$. If a likely alternative hypothesis is that an event shows a slow but steady increase or decrease in frequency over a period of nine years, construct a variate $X_i = 1, 2, 3, \dots, 9$ or alternatively $-4, -3, -2, \dots, +3, +4$ (making $\bar{X} = 0$), to represent the years. The average change in the f_i per year is estimated by the regression coefficient $\Sigma f_i x_i / \Sigma x_i^2$, where as usual $x_i = X_i - \bar{X}$. The value of χ^2 for testing this coefficient, against the null hypothesis that there is no change, is

$$\chi^2 = (\Sigma f_i x_i)^2 / f \Sigma x_i^2,$$

with 2 *df*.

Another example is found in an experiment designed to investigate various treatments for the control of cabbage loopers (insect larvae) (5). Each treatment was tested on four plots. Table 9.4.1 shows, for five of the treatments, the numbers of loopers counted on each plot. The objective of the analysis is to examine whether the treatments produced differences in the average number of loopers per plot.

TABLE 9.4.1
NUMBER OF LOOPERS ON 50 CABBAGE PLANTS IN A PLOT
(Four plots treated alike; five treatments)

Treatment	No. of Loopers Per Plot	Plot Total	Plot Mean	χ^2	<i>df</i> .
1	11, 4, 4, 5	24	6.00	5.67	3
2	6, 4, 3, 6	19	4.75	1.42	3
3	8, 6, 4, 11	29	7.25	3.69	3
4	14, 27, 8, 18	67	16.75	11.39	3
5	7, 4, 9, 14	34	8.50	6.24	3
Total		173		28.41	15

Since the sum of a number of independent Poisson variables also follows a Poisson distribution (section 8.14), we can compare the treatment totals by the Poisson variance test, provided we can adopt the assumption that the counts on plots treated alike follow the same Poisson distribution. To test this assumption, the χ^2 values for each treatment are computed in table 9.4.1 (second column from the right). Although only one of the five χ^2 values is significant at the 5% level, their total, 28.41, *df*. = 15, gives *P* of about 0.02. This finding invalidates the use of the Poisson variance test for the comparison of treatment totals. Some additional source of variation is present, which must be taken into account when investigating whether plot means differ from treatment to treatment. Problems of this type, which are common, are handled by the technique known as the analysis of variance. The analysis of these data is completed in example 10.3.3, p. 263.

Incidentally, the Poisson variance χ^2 for comparing the treatment totals would be computed as

$$\begin{aligned}\chi^2 &= \Sigma(f_i - \bar{f})^2 / \bar{f} \\ &= [(24)^2 + (19)^2 + \dots + (34)^2 - (173)^2/5] / 34.6 = 41.5,\end{aligned}$$

with 4 *d.f.* The high value of this χ^2 suggests that the variation between treatments is substantially greater than the variation within treatments—the point to be examined in the analysis of variance test.

EXAMPLE 9.4.1—In section 8.4, random numbers were used to draw 100 samples from the binomial $n = 5$, $p = 0.2$. The observed and expected frequencies (taken from table 8.4.1) are as follows

No. of Successes	0	1	2	3	4	5	Total
Observed frequency	32	44	17	6	1	0	100
Expected frequency	32.77	40.96	20.48	5.12	0.64	0.03	100.00

Compute χ^2 and test whether the deviations can be accounted for by sampling errors
Ans. $\chi^2 = 1.09$, *d.f.* = 3. *P* about 0.75. (Combine classes 3, 4, 5 before computing χ^2)

9.5—The χ^2 test when the expectations are small. The χ^2 test is a large-sample approximation, based on the assumption that the distributions of the observed numbers f_i (or y_i) in the classes are not far from normal. This assumption fails when some or all of the observed numbers are very small. Historically, the advice most often given was that the *expected* number in any class should not be less than 5, and that, if necessary, neighboring classes should be combined to meet this requirement. Later research, described in (6), showed that this restriction is too strict. Moreover, the combination of classes weakens the sensitivity of the χ^2 test.

We suggest that the χ^2 test is accurate enough if the smallest expectation is at least 1, and that classes be combined only to ensure this condition. This recommendation applies to the χ^2 tests of *single* classifications described in sections 9.2, 9.3, and 9.4. When counting the *d.f.* in χ^2 , the number of classes is the number after any necessary combinations have been made.

In more extreme cases it is possible to work out the exact distribution of χ^2 . The probability that f_i observations fall in the *i*th class is given by the *multinomial distribution*

$$\frac{n!}{f_1! f_2! \dots f_k!} p_1^{f_1} p_2^{f_2} \dots p_k^{f_k},$$

where the p_i are the probabilities specified by the null hypothesis. This distribution reduces to the binomial distribution when there are only two classes. This probability is evaluated, along with the value of χ^2 , for every possible set of f_i with $\Sigma f_i = n$.

When the expectations are equal (section 9.3), Chakravarti and Rao

7) have tabulated the exact 5% levels of χ^2 for samples in which $n = \sum f_i \leq 12$ and the number of classes, k , ≤ 100 . Our $\sum f_i$ is their T and our k is their f . Their tabulated criterion (in their table 1) is our $\sum f_i^2$, which is equivalent to χ^2 and quicker to compute.

EXAMPLE 9.5.1—When 5 dice were tossed 100 times, the observed and expected numbers of 2's out of 5 were as follows (data from example 1.9.8):

Number of 2's	f	F
5	2	0.013
4	3	0.322
3	3	3.214
2	18	16.075
1	42	40.188
0	32	40.188
Total	100	100.000

Applying the rule that the smallest expectation should be at least 1, we would combine classes 5, 4, 3. Verify that this gives $\chi^2 = 7.56$, $d.f. = 3$, P slightly above 0.05. Note that if we combined only the first two classes, this would give $\chi^2 = 66.45$, $d.f. = 4$.

9.6—Single classifications with estimated expectations. In sections 9.2 and 9.3, the null hypothesis specified the actual numerical values of the expectations in the classes. Often the null hypothesis gives these expectations in terms of one or more population parameters that must be estimated from the sample. This is so, for instance, in testing whether the observed frequencies of 0, 1, 2, . . . occurrences will fit the successive terms of a Poisson distribution. Unless the null hypothesis provides the value of μ , this must be estimated from the sample in order to calculate the expected frequencies. The estimate of μ is, of course, the sample mean.

The data of table 8.14.2, to which we have already fitted a Poisson distribution, serve as an example of the test of goodness of fit. The data and subsequent calculations appear in table 9.6.1. Having obtained the expected frequencies, we combine the last four classes (8 or more) so as to reach an expectation of at least 1. The deviations $(f - F)$ and the contributions $(f - F)^2/F$ to χ^2 are calculated as usual and given in the last two columns. We find $\chi^2 = 8.26$.

The only new step is the rule for counting the number of $d.f.$ in χ^2 :

$$d.f. = (\text{No. of classes}) - (\text{No. of estimated parameters}) - 1$$

In applying this rule, the number of classes is counted *after making* any combination of classes that is necessary because of small expectations. Each estimated parameter places one additional restriction on the sizes of the deviations $(f - F)$. The condition that $\sum (f - F) = 0$ also reduces the likely size of χ^2 . In this example the number of classes (after combining) is 9, and one parameter, μ , was estimated in fitting the

TABLE 9.6 1
 χ^2 TEST OF GOODNESS OF FIT OF THE POISSON DISTRIBUTION, APPLIED TO THE NUMBERS
 OF NOXIOUS WEED SEEDS FOUND IN 98 BATCHES

No. of Noxious Seeds	Observed Frequency (f)	Expected Frequency (F)	Observed - Expected ($f - F$)	Contribution to χ^2 : $\frac{(f - F)^2}{F}$
0	3	4.78	-1.78	0.66
1	17	14.44	+2.56	0.45
2	26	21.81	+4.19	0.80
3	16	21.96	-5.96	1.62
4	18	16.58	+1.42	0.12
5	9	10.02	-1.02	0.10
6	3	5.04	-2.04	0.83
7	5	2.18	+2.82	3.65
8	0	0.82	-0.20	0.03
9	1	0.27		
10	0	0.08		
11 or more	0	0.03		
Total	98	98.01	-0.01	8.26

distribution. Hence, there are $9 - 1 - 1 = 7$ *d.f.* The *P* value lies between 0.50 and 0.25. The fit is satisfactory.

Tests of this kind, in which we compare an observed frequency distribution with a theoretical distribution like the Poisson, the binomial, or the normal, are called *goodness of fit* tests. For the binomial, the *d.f.* are 2 less than the number of classes if *p* is estimated from the data, and 1 less than the number of classes if *p* is given in advance. With the normal, both parameters μ and σ are usually estimated, so that we subtract 3 from the number of classes.

You now have two methods of testing whether a sample follows the Poisson distribution: the goodness of fit test of this section and the variance test of section 9.3. If the members of the population actually follow Poisson distributions with *different* means, the variance test is more sensitive in detecting this than the goodness of fit test. The goodness of fit test is a general-purpose test, since *any* type of difference between the observed and expected numbers, if present in sufficient force, makes χ^2 large. But if something is known about the nature of the alternative hypothesis, we can often construct a different test that is more powerful for this type of alternative. The same remarks apply to the binomial distribution. A variance test for the binomial is given in section 9.8.

EXAMPLE 9.6 1 The numbers of tomato plants attacked by spotted wilt disease were counted in each of 160 areas of 9 plants (8). In all, 261 plants were diseased out of $9 \times 160 = 1440$ plants. A binomial distribution with $n = 9$, $p = 261/1440$, was fitted to the distribution of numbers of diseased plants out of 9. The observed and expected numbers are as follows

No. of Diseased Plants	0	1	2	3	4	5	6	7	Total
Observed frequency	36	48	38	23	10	3	1	1	160
Expected frequency	26.45	52.70	46.67	24.11	8.00	1.77	0.25	0.03	159.98

Perform the χ^2 goodness of fit test. Ans. $\chi^2 = 10.28$, with 4 *df.* after combining. $P < 0.05$.

EXAMPLE 9.6.2—In a series of trials a set of r successes, preceded and followed by a failure, is called a *run* of length r . Thus the series *FSFSSSF* contains one run of successes of length 1 and one of length 3. If the probability of a success is p at each trial, the probability of a run of length r may be shown to be $p^r q$. In 207 runs of diseased plants in a field, the frequency distribution of lengths of run was as follows:

Length of run	r	1	2	3	4	5	Total
Observed frequency	f_r	164	33	9	1	0	207

The estimate of p from these data is $\hat{p} = (T - N)/T$, where $N = \sum f_r = 207$ is the total number of runs and $T = \sum rf_r$ is the total number of successes in these runs. Estimate p ; fit the distribution, called the *geometric distribution*; and test the fit by χ^2 . Ans. $\chi^2 = 0.96$ with 2 *df.* $P > 0.50$. Note: the expression $(T - N)/T$, used for estimating p , is derived from a general method of estimation known as the method of maximum likelihood, and is not meant to be obvious. The expected frequency of runs of length r is $Np^{r-1}q$.

EXAMPLE 9.6.3—In table 3.4.1 (p. 71) a normal distribution was fitted to 511 means of samples of pig weight gains. Indicate how you would combine classes in making a goodness of fit test. How many *df.* does your χ^2 have? Ans. 17 *df.*

EXAMPLE 9.6.4—Apply the variance test for the Poisson distribution to the data in table 9.6.1. Ans. $\chi^2 = 105.3$ with 97 *df.* $P > 0.25$.

9.7—Two-way classifications. The $2 \times C$ contingency table. We come now to data classified by two different criteria. The simplest case (the 2×2 table), in which each classification has only two classes, was discussed in section 8.10. The next simplest case occurs when one classification has only two classes, the other having $C > 2$ classes. In the example in table 9.7.1, leprosy patients were classified at the start of an experiment according as to whether they exhibited little or much infiltration (a measure of a certain type of skin damage). They were also classified into five

TABLE 9.7.1
196 PATIENTS CLASSIFIED ACCORDING TO CHANGE IN HEALTH AND DEGREE OF INFILTRATION

Degree of Infiltration	Change in Health			Stationary	Worse	Total
	Improvement					
	Marked	Moderate	Slight			
Little	11	27	42	53	11	144
Much	7	15	16	13	1	52
Total	18	42	58	66	12	196

classes according to the change in their general health during a subsequent 48-week period of treatment (9). The patients did not all receive the same drugs, but since no differences in the effects of these drugs could be detected, the data were combined for this analysis. The table is called a 2×5 contingency table.

The question at issue is whether the change in health is related to the initial degree of infiltration. The χ^2 test extends naturally to $2 \times C$ tables. The overall proportion of patients with little infiltration is $144/196$. On the null hypothesis of no relationship between degree of infiltration and change in health, we expect to find $(18)(144)/196 = 13.22$ patients with little infiltration and marked improvement, as against 11 observed. As before, the rule for finding an expected number is (row total)(column total)/(grand total). The expected numbers F and the deviations $(f - F)$ are shown in table 9.7.2. Note that only four expected numbers need be calculated: the rest can be found by subtraction.

TABLE 9.7.2
EXPECTED NUMBERS AND DEVIATIONS CALCULATED FROM TABLE 9.7.1

Degree of Infiltration	Change in Health					Total
	Improvement		Stationary	Worse		
	Marked	Moderate			Slight	
	<i>Expected numbers, F</i>					
Little	13.22	30.86	42.61	48.49	8.82	144.00
Much	4.78	11.14	15.39	17.51	3.18	52.00
Total	18.00	42.00	58.00	66.00	12.00	196.00
	<i>Deviations, (f - F)</i>					
Little	-2.22	-3.86	-0.61	+4.51	+2.18	0.00
Much	+2.22	+3.86	+0.61	-4.51	-2.18	0.00

The value of χ^2 is

$$\chi^2 = \sum (f - F)^2 / F$$

$$= (-2.22)^2 / 13.22 + (+2.22)^2 / 4.78 + \dots + (-2.18)^2 / 3.18 = 6.87,$$

taken over the ten cells in the table. The number of *d.f.* is $(R - 1)(C - 1)$, where R , C are the numbers of rows and columns, respectively. In this example $R = 2$, $C = 5$ and we have 4 *d.f.* This rule for *d.f.* is in line with the fact that when four of the deviations in a row are known, all the rest can be found. With $\chi^2 = 6.87$, *d.f.* = 4, the probability lies between 0.25 and 0.10.

Although this test has not rejected the null hypothesis, the deviations show a systematic pattern. In the "much infiltration" class, the observed numbers are higher than expected for patients showing any degree of improvement, and lower than expected for patients classified as sta-

tionary or worse. The reverse is, of course, true for the "little infiltration" class. Contrary to the null hypothesis, these deviations suggest that patients with much infiltration progressed on the whole better than those with little infiltration. This suggestion will be studied further in section 9.10.

9.8—The variance test for homogeneity of the binomial distribution.

In the preceding example we obtained a $2 \times C$ contingency table because the data were classified into 2 classes by one criterion and into C classes by a second criterion. Alternatively, we may have recorded some binomial variate $p_i = a_i/n_i$ in each of C independent samples, where i goes from 1 to C and n_i is the size of the i th sample. The objective now is to examine whether the true p_i vary from sample to sample. Data of this type occur very frequently.

A quicker method of computing χ^2 which is particularly appropriate in this situation was devised by Snedecor and Irwin (10). It will be illustrated by the preceding example. Think of the columns in table 9.8.1 as representing $C = 5$ samples.

TABLE 9.8.1
ALTERNATIVE CALCULATION OF χ^2 FOR THE DATA IN TABLE 9.7.1

Degree of Infiltration	Improvement					Total
	Marked	Moderate	Slight	Stationary	Worse	
Little	11	27	42	53	11	144
Much (a_i)	7	15	16	13	1	52 (A)
Total (n_i)	18	42	58	66	12	196 (N)
$p_i = a_i/n_i$	0.3889	0.3571	0.2759	0.1970	0.0833	0.26531 (\bar{p})

First calculate the proportion $p_i = a_i/n_i$ of "much infiltration" patients in each column, and the corresponding overall proportion $\bar{p} = A/N = 52/196 = 0.26531$. Then,

$$\begin{aligned}
 \chi^2 &= (\sum p_i a_i - \bar{p} A) / \bar{p} \bar{q} \\
 &= [(0.3889)(7) + \dots + (0.0833)(1) \\
 &\quad - (0.26531)(52)] / (0.26531)(0.73469) \\
 &= 6.88,
 \end{aligned} \tag{9.8.1}$$

as before, with 4 *d.f.*

If p_i is the variable of interest, you will want to calculate these values anyway in order to examine the results. Extra decimals should be carried to ensure accuracy in computing χ^2 , particularly when the a_i are large. The computations are a little simpler when the p_i are derived from the row with the *smaller* numbers.

This formula for χ^2 can be written, alternatively,

$$\chi^2 = \sum n_i (p_i - \bar{p})^2 / \bar{p} \bar{q} \tag{9.8.2}$$

If the binomial estimates p_i are all based on the same sample size n , χ^2 becomes

$$\chi^2 = \sum_{i=1}^C (p_i - \bar{p})^2 / (\bar{p}\bar{q}/n) = (C-1)s_p^2 / (\bar{p}\bar{q}/n) \quad (9.8.3)$$

In this form, χ^2 is essentially a comparison of the observed variance s_p^2 among the p_i with the variance $\bar{p}\bar{q}/n$ that the p_i would have if they were independent samples from the same binomial distribution. The same interpretation can be shown to apply to expression (9.8.2) for χ^2 . A high value of χ^2 denotes that the true proportions differ from sample to sample.

This test, sometimes called the variance test for homogeneity of the binomial distribution, has many applications. Different investigators may have estimated the same proportion in different samples, and we wish to test whether the estimates agree, apart from sampling errors. In a study of an attribute in human families, where each sample is a family, a high value of χ^2 indicates that members of the same family tend to be alike with regard to this attribute.

When some of the sample sizes n_i are small, some of the expectations $n_i\bar{p}$ and $n_i\bar{q}$ will be small. The χ^2 test can still be used with some expectations as low as 1, provided that most of the expectations (say 4 out of 5) are substantially larger. (Recent results [11] suggest that this advice is conservative.) In some genetic and family studies, *all* the n_i are small. For this case a good approximation to the significance levels of the exact χ^2 distribution has been given by Haldane (12), though the computations are laborious. When χ^2 has more than 30 *d.f.* and the n_i are all equal ($=n$) the exact χ^2 is approximately normally distributed with

$$\text{Mean} = (C-1)N/(N-1)$$

$$\begin{aligned} \text{Variance} &= 2(C-1) \left(\frac{n-1}{2} \right) \frac{N^4}{(N-1)^2(N-2)(N-3)} \left[1 - \frac{(N-1)}{A(N-A)} \right] \\ &\doteq 2(C-1) \left(\frac{n-1}{n} \right) \left[1 + \frac{1}{N} \left(7 - \frac{1}{\bar{p}\bar{q}} \right) \right], \end{aligned}$$

where C is the number of samples and $N = Cn$.

When the p_i vary from column to column, as indicated by a high value of χ^2 , the binomial formula $\sqrt{(\bar{p}\bar{q}/N)}$ underestimates the standard error of the overall proportion \bar{p} for the combined sample. A more nearly correct formula (section 17.5) for the standard error of \bar{p} in this situation is

$$s.e.(\bar{p}) = \frac{1}{\bar{n}} \sqrt{(\sum a_i^2 - 2\bar{p}\sum a_i n_i + \bar{p}^2 \sum n_i^2) / C(C-1)}, \quad (9.8.4)$$

where C is the number of samples and

$$p_i = \frac{a_i}{n_i} : \bar{p} = \frac{\sum a_i}{\sum n_i} : N = \sum n_i : \bar{n} = \frac{N}{C}$$

EXAMPLE 9.8.1—Ten samples of 5 mice from the same laboratory were injected with the same dose of *bact. typhi. murium* (13). The numbers of mice dying (out of 5) were as follows: 3, 1, 5, 5, 3, 2, 4, 2, 3, 5. Test whether the proportion dying can be regarded as constant from sample to sample. Ans. $\chi^2 = 18.1$, $d.f. = 9$. $P < 0.05$. Since the death rate is found so often to vary within the same laboratory, a standard agent is usually tested along with each new agent, because comparisons made over time cannot be trusted.

EXAMPLE 9.8.2—Uniform doses of *Danysz bacillus* were injected into rats, the sizes of the samples being dictated by the numbers of animals available at the dates of injection. These sizes, the numbers of surviving rats, and the proportion surviving, are as follows:

Number in sample	40	12	22	11	37	20
Number surviving	9	2	3	1	2	3
Proportion surviving	0.2250	0.1667	0.1364	0.0909	0.0541	0.1500

Test the null hypothesis that the probability of survival is the same in all samples. Ans. $\chi^2 = 4.97$, $d.f. = 5$, $P = 0.43$.

EXAMPLE 9.8.3—In another test with four samples of inoculated rats, χ^2 was 6.69, $P = 0.086$. Combine the values of χ^2 for the two tests. Ans. $\chi^2 = 11.66$, $d.f. = 8$, $P = 0.17$.

EXAMPLE 9.8.4—Burnett (14) tried the effect of five storage locations on the viability of seed corn. In the kitchen garret, 111 kernels germinated among 120 tested; in a closed toolshed, 55 out of 60; in an open toolshed, 55 out of 60; outdoors, 41 out of 48; and in a dry garret, 50 out of 60. Calculate $\chi^2 = 5.09$, $d.f. = 4$, $P = 28\%$.

EXAMPLE 9.8.5—In 13 families in Baltimore, the numbers of persons (n_i) and the numbers (a_i) who had consulted a doctor during the previous 12 months were as follows: 7, 0; 6, 0; 5, 2; 5, 5; 4, 1; 4, 2; 4, 2; 4, 2; 4, 0; 4, 0; 4, 4; 4, 0; 4, 0. Compute the overall percentage who had consulted a doctor and the standard error of the percentage. Note: One would expect the proportion who had seen a doctor to vary from family to family. Verify this by finding $\chi^2 = 35.6$, $d.f. = 12$, $P < 0.005$. Consequently, formula 9.8.4 is used to estimate the *s.e.* of \bar{p} . Ans. Percentage = $100\bar{p} = 30.5\%$, *s.e.* = 10.5% . (These data were selected from a large sample for illustration.)

9.9—Further examination of the data. When the initial χ^2 test shows a significant value, the remarks made in section 9.2 about further examination of the data apply here also. Subsequent tests are made that may help to explain the high value of χ^2 . Frequently, as already remarked, the investigator proceeds at once to these tests, omitting the initial χ^2 test as not informative.

Decker and Andre (15) investigated the effect of a short, sudden exposure to cold on the adult chinch bug. Since experimental insects had to be gathered in the field, the degree of heterogeneity in the insects was unknown, and the investigators faced the problem as to whether they could reproduce their results. Ten adult bugs were placed in each of 50 tubes and exposed for 15 minutes at -8°C . For this illustration the counts of the numbers dead in the individual tubes were combined at random into 5 lots of 10 tubes each; that is, into lots of 100 chinch bugs. The numbers dead were 14, 14, 23, 17, and 20 insects. From these data, $\chi^2 = 4.22$, $d.f. = 4$, $P = 0.39$. The results are in accord with the hypothesis that every adult bug was subject to the same chance of being killed by the exposure.

In a second sample of 500 adults, handled in the same manner except that they were exposed at $-9^{\circ}\text{C}.$, the numbers dead in groups of 100 were 38, 30, 30, 40, 27. The χ^2 value of 5.79 again verifies the technique, showing only sampling variation from the estimated mortality of 33%.

The gratifying uniformity in the results leads one to place some confidence in the surprising finding that the death rates at $-8^{\circ}\text{C}.$ and $-9^{\circ}\text{C}.$ were markedly different. The total numbers dead in the two samples of 500 were 88 and 165. The result, $\chi^2 = 31.37$ with $d.f. = 1$, P less than 0.0002, provides convincing evidence that a rise in mortality with the lowering of temperature from $-8^{\circ}\text{C}.$ to $-9^{\circ}\text{C}.$ is a characteristic of the population, not merely an accident of sampling.

The ease of applying a test of experimental technique makes its use almost a routine procedure except in highly standardized processes. It is necessary merely to collect the data in several small groups, chosen with regard to the types of experimental variation thought likely to be present, instead of in one mass. The additional information may modify conclusions and subsequent procedures profoundly.

In this example the sum of the three values of χ^2 is $4.22 + 5.79 + 31.37 = 41.38$, with 9 $d.f.$ If the initial χ^2 is calculated from the 2×10 contingency table formed by the complete data, its value is also found to be 41.38, with 9 $d.f.$ This agreement between the two values is a fluke, which does not hold generally in $2 \times C$ tables. For $2 \times C$ and $R \times C$ tables, a method of computing the component parts so that they add to the initial total χ^2 is available (16). In these data this method amounts to using the same denominator $\bar{p}\bar{q} = (0.253)(0.747)$, calculated from the *total* mortality, in finding all χ^2 values. Instead, for the 4 $d.f.$ χ^2 at $-8^{\circ}\text{C}.$ we used $\bar{p}\bar{q} = (0.176)(0.824)$, appropriate to that part of the data, and at $-9^{\circ}\text{C}.$ we used $\bar{p}\bar{q} = (0.330)(0.670)$. The additive χ^2 values give $3.24 + 6.77 + 31.37 = 41.38$. However, when it has been shown that the mortality differs at $-8^{\circ}\text{C}.$ and $-9^{\circ}\text{C}.$, use of a pooled \bar{p} for the individual homogeneity tests at $-8^{\circ}\text{C}.$ and $-9^{\circ}\text{C}.$ is invalid. The non-additive method is recommended, except in a quick preliminary look at the data.

9.10—Ordered classifications. In the leprosy example of section 9.7, the classes (marked improvement, moderate improvement, slight improvement, stationary, worse) are an example of an *ordered classification*. Such classifications are common in the study of human behavior and preferences, and more generally whenever different degrees of some phenomenon can be recognized. The problem of utilizing the knowledge that we possess about this ordering has attracted considerable attention in recent years.

With a single classification of Poisson variables, the ordering might lead us to expect that if the null hypothesis $\mu_1 = \mu$ does not hold, an alternative $\mu_1 \leq \mu_2 \leq \mu_3 \leq \dots$ should hold, where the subscripts represent the order. For instance, if working conditions in a factory have been classified as Excellent, Good, Fair, we might expect that if the number of defective articles per worker varies with working conditions, the order should

be $\mu_1 \leq \mu_2 \leq \mu_3$. Similarly, with ordered columns in a $2 \times C$ contingency table, the alternative $p_1 \leq p_2 \leq p_3$ might be expected. χ^2 tests designed to detect this type of alternative have been developed by Bartholomew (17). The computations are quite simple.

Another approach, used by numerous workers (9), (18), (19), is to attach a score to each class so that an ordered scale is created. To illustrate from the leprosy example, we assigned scores of 3, 2, 1, respectively, to the Marked, Moderate, and Slight Improvement classes, 0 to the Stationary class, and -1 to the Worse class. These scores are based on the judgment that the five classes constructed by the expert represent equal gradations on a continuous scale. We considered giving a score of $+4$ to the Marked Improvement class and -2 to the Worse class, since the expert seemed to examine a patient at greater length before assigning him to one of these extreme classes, but rejected this since our impression may have been erroneous.

Having assigned the scores we may think of the leprosy data as consisting of two independent samples of 144 and 52 patients, respectively. (See table 9.10.1.) For each patient we have a discrete measure X of his change in health, where X takes only the values 3, 2, 1, 0, -1 . We can estimate the average change in health for each sample, with its standard error, and can test the null hypothesis that this average change is the same in the two populations. For this test we use the ordinary two-sample t -test as applied to grouped data. The calculations appear in table 9.10.1. On the X scale the average change in health is $+1.269$ for patients with much infiltration and $+0.819$ for those with little infiltration. The difference, \bar{D} , is 0.450 , with standard error ± 0.172 (194 d.f.), computed in the usual way. The value of t is $0.450/0.172 = 2.616$, with $P < 0.01$. Contrary to the initial χ^2 test, this test reveals a significantly greater amount of progress for the patients with much infiltration.

The assignment of scores is appropriate when (i) the phenomenon in question is one that could be measured on a continuous scale if the instruments of measurement were good enough, and (ii) the ordered classification can be regarded as a kind of grouping of this continuous scale, or as an attempt to approximate the continuous scale by a cruder scale that is the best we can do in the present state of knowledge. The process is similar to that which occurs in many surveys. The householder is shown five specific income classes and asked to indicate the class within which his income falls, without naming his actual income. Some householders name an incorrect class, just as an expert makes some mistakes in classification when this is difficult.

The advantage in assigning scores is that the more flexible and powerful methods of analysis that have been developed for continuous variables become available. One can begin to think of the sizes of the average differences between different groups in a study, and compare the difference between groups A and B with that between groups E and F . Regressions of the group means \bar{X} on a further variable Z can be worked

TABLE 9.10.1
ANALYSIS OF THE LEPROSY DATA BY ASSIGNED SCORES
(Data with assigned scores)

Change in Health	Infiltration	
	Little	Much
X	<i>No. of patients</i>	
	f	f
3	11	7
2	27	15
1	42	16
0	53	13
-1	11	1
Total: Σf	144	52

(Computations)

	<i>Little</i>	<i>Much</i>
ΣfX	118	66
$\bar{X} = \Sigma fX / \Sigma f$	0.819	1.269
ΣfX^2	260	140
$(\Sigma fX)^2 / \Sigma f$	96.7	83.8

Σfx^2	163.3	56.2
$d.f.$	143	51
s^2	1.142	1.102

Pooled s^2 1.131

$$s_D^2 = (1.131) \left(\frac{1}{144} + \frac{1}{52} \right) = 0.0296$$

$$s_D = 0.172$$

$$t = \frac{\bar{D}}{s_D} = \frac{1.269 - 0.819}{0.172} = 2.616$$

$$d.f. = 194, \quad P < 0.01$$

out. The relative variability of different groups can be examined by computing s for each group.

This approach assumes that the standard methods of analysis of continuous variables, like the t -test, can be used with an X variable that is discrete and takes only a few values. As noted in section 5.8 on scales with limited values, the standard methods appear to work well enough for practical use. However, heterogeneity of variance and correlation between s^2 and \bar{X} are more frequently encountered because of the discrete scale. If most of the patients in a group show marked improvement, most of their X 's will be 3, and s^2 will be small. Pooling of variances should not be undertaken without examining the individual s^2 . In the leprosy example the two s^2 were 1.142 and 1.102 (table 9.10.1), and this difficulty was not present.

The chief objection to the assignment of scores is that the method is more or less arbitrary. Two investigators may assign different scores to the same set of data. In our experience, however, moderate differences between two scoring systems seldom produce marked differences in the conclusions drawn from the analysis. In the leprosy example, the alternative scores 4, 2, 1, 0, -2 give $t = 2.549$ as against $t = 2.616$ in the analysis in table 9.10.1. Some classifications present particular difficulty. If the degrees of injury to persons in accidents are recorded as slight, moderate, severe, disabling, and fatal, there seems no entirely satisfactory way of placing the last two classes on the same scale as the first three.

Several alternative principles have been used to construct scores. In studies of different populations of school children, K. Pearson (20) assumed that the underlying continuous variate was normally distributed in a standard population of school children. If the classes are regarded as a grouping of this normal distribution, the class boundaries for the normal variate are easily found. The score assigned to a class is the mean of the normal variate within the class. A related approach due to Bross (21) also uses a standard population but does not assume normality. The score (*ridit*) given to a class is the relative frequency up to the midpoint of that class in the standard population. When the experimental treatments are different doses of a toxic or protective agent in biological assay, Ipsen (22) shows how to assign scores so that the resulting variate has a linear regression on some chosen function of the dose, the ratio of the variance due to regression to the total variance being maximized. Fisher (23) assigns scores so as to maximize the F -ratio of treatments to experimental error as defined in section 10.5. The *maximin* method of Abelson and Tukey (24), maximizes the square of the correlation coefficient r^2 between the assigned scores and the set of true scores, consistent with the investigator's knowledge about the ordering of the classes, that gives a minimum correlation with the assigned scores. This approach, like Bartholomew's, avoids any arbitrary assumptions about the nature of the true scale.

EXAMPLE 9 10 1—In the leprosy data, verify the value of $t = 2.549$ quoted for the scoring 4, 2, 1, 0, -2

9.11—Test for a linear trend in proportions. When interest is centered on the proportions p_i in a $2 \times C$ contingency table, there is another way of viewing the data. Table 9.11.1 shows the leprosy data with the assigned scores X_i , but in this case the variable that we analyze is p_i , the proportion of patients with much infiltration. The contention now is that if these patients have fared better than patients with little infiltration, the values of p_i should increase as we move from the Worse class ($X = -1$) towards the Marked Improvement class ($X = 3$).

If this is so, the regression coefficient of p_i on X_i should be a good test criterion. On the null hypothesis (no relation between p_i and X_i) each p_i is distributed about the same mean, estimated by \bar{p} , with variance $\bar{p}\bar{q}/n_i$. The regression coefficient b is calculated as usual, except that each p_i must be weighted by the reciprocal of the sample size n_i on which it is

TABLE 9.11.1
TESTING A LINEAR REGRESSION OF p_i ON THE SCORE (LEPROSY DATA)

Degree of Infiltration	Improvement			Stationary	Worse	Total
	Marked	Moderate	Slight			
Little	11	27	42	53	11	144
Much (a_i)	7	15	16	13	1	52
Total (n_i)	18	42	58	66	12	196 (N)
$p_i = a_i/n_i$	0.3889	0.3571	0.2759	0.1970	0.0833	0.2653 (\bar{p})
Score X_i	3	2	1	0	-1	

based. The numerator and denominator of b are computed as follows:

$$\begin{aligned}
 \text{Num.} &= \sum n_i(p_i - \bar{p})(X_i - \bar{X}) \\
 &= \sum n_i p_i X_i - (\sum n_i p_i)(\sum n_i X_i) / \sum n_i \\
 &= \sum a_i X_i - (\sum a_i)(\sum n_i X_i) / N \\
 &= 66 - (52)(184) / 196 = 66 - 48.82 = 17.18 \\
 \text{Den.} &= \sum n_i X_i^2 - (\sum n_i X_i)^2 / N \\
 &= 400 - (184)^2 / 196 = 400 - 172.8 = 227.2
 \end{aligned}$$

This gives $b = 17.18 / 227.2 = 0.0756$. Its standard error is

$$s_b = \sqrt{(\bar{p}\bar{q} / \text{Den.})} = \sqrt{\{(0.2653)(0.7347) / (227.2)\}} = 0.0293$$

The normal deviate for testing the null hypothesis $\beta = 0$ is

$$Z = b/s_b = 0.0756 / 0.0293 = 2.580. \quad P = 0.0098.$$

Although it is not obvious at first sight, Yates (18) showed that this regression test is essentially the same as the t -test in section 9.10 of the difference between the mean scores in the Little and Much infiltration classes. In this example the regression test gave $Z = 2.580$ while the t -test gave $t = 2.616$ (194 d.f.). The difference in results arises because the two approaches use slightly different large-sample approximations to the exact distributions of Z and t with these discrete data.

EXAMPLE 9.11.1 Armitage (19) quotes the following data by Holmes and Williams for the relation in children between size of tonsils and the proportion of children who are carriers of *streptococcus pyogenes* in the nose

Types of Children	X = Score Given to Size of Tonsils			Total Children
	0	1	2	
Carriers (a_i)	19	29	24	72 (A)
Non-carriers	497	560	269	1326
Total (n_i)	516	589	293	1398 (N)
Carrier-rate (p_i)	0.0368	0.0492	0.0819	0.051502 (\bar{p})

Calculate: (i) the normal deviate Z for testing the linear regression of the proportion of carriers on size of tonsils, (ii) the value of t for comparing the difference between the mean size of tonsils in carriers and non-carriers. Ans. (i) $Z = 2.681$, (ii) $t = 2.686$, with 1396 *d.f.*

EXAMPLE 9.11.2—When the regression of p_i on X_i is used as a test criterion, it is of interest to examine whether the regression is linear. Armitage (19) shows that this can be done by first computing $\chi^2 = \sum n_i(p_i - \bar{p})^2/\bar{p}\bar{q} = \{\sum a_i p_i - A^2/N\}/\bar{p}\bar{q}$. This χ^2 , with $(C - 1)$ *d.f.*, measures the total variation among the C values of p_i . The χ^2 for linear regression, with 1 *d.f.*, is found by squaring Z , since the square of a normal deviate has a χ^2 distribution with 1 *d.f.* The difference, $\chi_{(C-1)}^2 - \chi_1^2$, is a χ^2 with $(C - 2)$ *d.f.* for testing the deviations of the p_i from their linear regression on the X_i . Compute this χ^2 for the data in example 9.11.1. Ans. The total χ^2 is 7.85 with 2 *d.f.*, while Z^2 is 7.19 with 1 *d.f.* Thus the χ^2 for the deviations is 0.66 with 1 *d.f.*, in agreement with the hypothesis of linearity.

9.12—Heterogeneity χ^2 in testing Mendelian ratios. It is often advisable to collect data in several small samples rather than in a single large one. An example is furnished by some experiments on chlorophyll inheritance in maize (1), reported in table 9.12.1. The series consisted of 11 samples of progenies of heterozygous green plants, self-fertilized, segregating into dominant green plants and recessive yellow plants. The hypothetical ratio is 3 green to 1 yellow. We shall study the proportion of yellow—theoretically 1/4.

TABLE 9.12.1
NUMBER OF YELLOW SEEDLINGS IN 11 SAMPLES OF MAIZE

No. in Sample	No. Yellow	Proportion Yellow
n_i	a_i	$p_i = a_i/n_i$
122	24	0.1967
149	39	0.2617
86	18	0.2093
55	13	0.2364
71	17	0.2394
179	38	0.2123
150	30	0.2000
36	9	0.2500
91	21	0.2308
53	14	0.2642
111	26	0.2342
$N = 1103$	$A = 249$	$\bar{p} = 0.22575$

Heterogeneity χ^2 (10 *d.f.*)

$$\chi^2 = (\sum a_i p_i - A\bar{p})/\bar{p}\bar{q} = (0.5779)/(0.2258)(0.7742) = 3.31$$

Pooled χ^2 (1 *d.f.*)

$$\chi_c^2 = (|A - N\bar{p}| - \frac{1}{2})^2/N\bar{p}\bar{q} \\ = (|249 - 275.75| - \frac{1}{2})^2/(1103)(0.25)(0.75) = 3.33$$

The data may fail to satisfy the simple Mendelian hypothesis in two ways. First, there may be real differences among the p_i (proportion of yellow) in different samples. This finding points to some additional source

of variability that must be explained before the data can be used as a crucial test of the Mendelian ratio. Second, the p_i may agree with one another (apart from sampling errors) but their overall proportion \bar{p} may disagree with the Mendelian proportion p . The reason may be linkage or crossing-over, or differential robustness in the dominant and recessive plants.

The first point is examined by applying to the p_i the variance test for homogeneity of the binomial distribution (section 9.8). The value of χ^2 shown under table 9.12.1, is 3.31, with 10 *d.f.*, P about 0.97. The test gives no reason to suspect real differences among the p_i . We therefore pool the samples and compare the overall ratio, $\bar{p} = 0.22575$, with the hypothetical $p = 0.25$, by the χ^2 test for a binomial proportion (section 8.8). We find χ^2 (corrected for continuity) = 3.33, P about 0.07. There is a hint of a deficiency of the recessive yellows.

In showing the relation between these two tests, the following algebraic identity is of interest:

$$\frac{\sum n_i(p_i - p)^2}{pq} = \frac{(\sum n_i)(\bar{p} - p)^2}{pq} + \frac{\sum n_i(p_i - \bar{p})^2}{pq} \quad (9.12.1)$$

The quantity $n_i(p_i - p)^2/pq$ measures the discrepancy between the observed p_i in the i th sample and the theoretical value p . If the null hypothesis is true, this quantity is distributed as χ^2 with 1 *d.f.* and the sum of these quantities over the C samples (left side of equation 9.12.1) is distributed as χ^2 with C *d.f.* The first term on the right of (9.12.1) compares the pooled ratio \bar{p} with p , and is distributed as χ^2 with 1 *d.f.* The second term on the right measures the deviations of the p_i from their own pooled mean \bar{p} , and is distributed as χ^2 with $(C - 1)$ *d.f.* To sum up, the total χ^2 on the left, with C *d.f.*, splits into a χ^2 with 1 *d.f.* which compares the pooled sample \bar{p} and the theoretical p , and a heterogeneity χ^2 , with $(C - 1)$ *d.f.*, which compares the p_i among themselves. These χ^2 distributions are of course followed only approximately unless the n_i are large.

In practice, this additive feature is less useful. Unless the pooled sample is large, a correction for continuity in the 1 *d.f.* for the pooled χ^2 is advisable. This destroys the additivity. Secondly, the expression for the heterogeneity χ^2 assumes that the theoretical ratio p applies in these data. If there is doubt on this point, the heterogeneity χ^2 should be calculated, as in table 9.12.1, with $\bar{p}q$ in the denominator instead of pq . In this form the heterogeneity χ^2 involves no assumption that $\bar{p} = p$ (apart from sampling errors).

EXAMPLE 9.12.1—From a population expected to segregate 1:1, four samples with the following ratios were drawn, 47:33, 40:26, 30:42, 24:34. Note the discrepancies among the sample ratios. Although the pooled χ^2 does not indicate any unusual departure from the theoretical ratio, you will find a large heterogeneity χ^2 equal to 9.01, $P = 0.03$, for which some explanation should be sought.

EXAMPLE 9 12 2—Fisher (25) applied χ^2 tests to the experiments conducted by Mendel in 1863 to test different aspects of his theory, as follows

Experiment	χ^2	df
Trifactorial	8.94	17
Bifactorial	2.81	8
Gametic ratios	3.67	15
Repeated 2 × 1 test	0.13	1

Show that in random sampling the probability of obtaining a total χ^2 lower than that observed is less than 0.005 (use the χ^2 table). More accurately, the probability is less than 1 in 2000. Thus the agreement of the results with Mendel's laws looks too good to be true. Fisher gives an interesting discussion of possible reasons.

9.13—The $R \times C$ table. If each member of a sample is classified by one characteristic into R classes, and by a second characteristic into C classes, the data may be presented in a table with R rows and C columns. The entry in any of the RC cells is the number of members of the sample falling into that cell. Strand and Jessen (26) classified a random sample of farms in Audubon County, Iowa, into three classes (Owned, Rented, Mixed), according to the tenure status and into three classes (I, II, III), according to the level of the soil fertility (table 9 13 1).

TABLE 9 13 1
NUMBERS OF FARMS ON THREE SOIL FERTILITY GROUPS IN AUDUBON COUNTY, IOWA,
CLASSIFIED ACCORDING TO TENURE

Soil		Owned	Rented	Mixed	Total
I	f	36	67	49	152
	F	36.75	62.92	52.33	
	$f - F$	-0.75	4.08	-3.33	
II	f	31	60	49	140
	F	33.85	57.95	48.20	
	$f - F$	-2.85	2.05	0.80	
III	f	58	87	80	225
	F	54.40	93.13	77.47	
	$f - F$	3.60	-6.13	2.53	
Total		125	214	178	517

$$\chi^2 = \sum \frac{(f - F)^2}{F} = \frac{(-0.75)^2}{36.75} + \frac{(2.53)^2}{77.47} = 1.54 \quad df = (R - 1)(C - 1) = 4$$

Before drawing conclusions about the border totals for tenure status, this question is asked: Are the relative numbers of Owned, Rented, and Mixed farms in this county the same at the three levels of soil fertility?

This question might alternatively be phrased Is the distribution of the soil fertility levels the same for Owned, Rented, and Mixed farms? (If a little reflection does not make it clear that these two questions are equivalent, see example 9 13 1) Sometimes the question is put more succinctly as Is tenure status *independent* of fertility level?

The χ^2 test for the $2 \times C$ table extends naturally to this situation. As before,

$$\chi^2 = \Sigma(f - F)^2/F,$$

where f is the observed frequency in any cell and F the frequency expected if the null hypothesis of independence holds

As before, the expected frequency for any cell is computed from the border totals in the corresponding row and column

$$\begin{aligned} F &= \frac{(\text{row total})(\text{column total})}{n} \\ &= \frac{\text{row total}}{n} (\text{column total}) \end{aligned}$$

Examples For the first row,

$$\begin{aligned} \frac{\text{row total}}{n} &= \frac{152}{517} = 0.29400 \\ F_1 &= (0.29400)(125) = 36.75 \\ F_2 &= (0.29400)(214) = 62.92 \\ F_3 &= (0.29400)(178) = 52.33 \end{aligned}$$

This procedure makes the computation easy with a calculating machine. For verification, notice that (i) the sum of the F in any row or column is equal to the observed total, and consequently (ii) the sum of the deviations in each row and in each column is zero

The facts just stated dictate the number of degrees of freedom. One is free to put $R - 1$ expected frequencies in a column, but the remaining cell is then fixed as the column total minus the sum of the $R - 1$ values of F . Similarly, when we have inserted expected frequencies in this way in $(C - 1)$ columns, the expected frequencies in the last column are fixed. Therefore $df = (R - 1)(C - 1)$.

The calculation of χ^2 is given in the table. Since $P > 0.8$, the null hypothesis is not rejected. If you do not need to examine the contribution of the individual cells of χ^2 , up to half the time in computation can be saved by a shortcut devised by P. H. Leslie (27). This is especially useful if many tables are to be calculated.

When χ^2 is significant, the next step is to study the nature of the departure from independence in more detail. Examination of the cells in which the contribution to χ^2 is greatest, taking note of the signs of the deviations ($f - F$), furnishes clues, but these are hard to interpret because the deviations in different cells are correlated. Computation of the per-

centage distribution of the row classification within each column, followed by a scrutiny of the changes from column to column, may be more informative. Further χ^2 tests may help. For instance, if the percentage distribution of the row classification appears the same in two columns, a χ^2 test for these two columns may confirm this. The two columns can then be combined for comparison with other columns. Examples 9.13.2, 3, 4, 5 illustrate this approach.

EXAMPLE 9.13.1—Show that if the expected distribution of the column classification is the same in every row, then the expected distribution of the row classification is the same in every column. For the i th row, let $F_{i1}, F_{i2}, \dots, F_{ic}$ be the expected numbers in the respective columns. Let $F_{i2} = a_2 F_{i1}, F_{i3} = a_3 F_{i1}, \dots, F_{ic} = a_c F_{i1}$. Then the numbers a_2, a_3, \dots, a_c must be the same in every row, since the expected distribution of the column classification is the same in every row. Now the expected row distribution in the first column is $F_{11}, F_{21}, \dots, F_{R1}$. In the second column it is $F_{12} = a_2 F_{11}, F_{22} = a_2 F_{21}, \dots, F_{R2} = a_2 F_{R1}$. Since a_2 is a constant multiplier, this is the same distribution as in the first column, and similarly for any other column.

EXAMPLE 9.13.2—In a study of the relation between blood type and disease, large samples of patients with peptic ulcer, patients with gastric cancer, and control persons free from these diseases were classified as to blood type (O, A, B, AB). In this example, the relatively small numbers of AB patients were omitted for simplicity. The observed numbers are as follows:

Blood Type	Peptic Ulcer	Gastric Cancer	Controls	Totals
<i>O</i>	983	383	2892	4528
<i>A</i>	679	416	2625	3720
<i>B</i>	134	84	570	788
Totals	1796	883	6087	8766

Compute χ^2 to test the null hypothesis that the distribution of blood types is the same for the three samples. Ans. $\chi^2 = 40.54, 4 \text{ df}$ P very small.

EXAMPLE 9.13.3—To examine this question further, compute the percentage distribution of blood types for each sample, as shown below.

Blood Type	Peptic Ulcer	Gastric Cancer	Controls
<i>O</i>	54.7	43.4	47.5
<i>A</i>	37.8	47.1	43.1
<i>B</i>	7.5	9.5	9.4
Totals	100.0	100.0	100.0

This suggests (i) there is little difference between the blood type distributions for gastric cancer patients and controls, (ii) peptic ulcer patients differ principally in having an excess of patients of type O . Going back to the frequencies in example 9.13.2, test the hypothesis that the blood type distribution is the same for gastric cancer patients and controls. Ans. $\chi^2 = 5.64 (2 \text{ df})$. P about 0.06.

EXAMPLE 9.13.4—Combine the gastric cancer and control samples. Test (i) whether the distribution of A and B types is the same in this combined sample as in the peptic ulcer sample (omit the O types). Ans. $\chi^2 = 0.68 (1 \text{ df})$ $P > 0.7$. (ii) Test whether proportion

of O types versus $A + B$ types is the same for the combined sample as for the gastric cancer samples. Ans. $\chi^2 = 34.29$ (1 *d.f.*). P very small. To sum up, the high value of the original 4 *d.f.* χ^2 is due primarily to an excess of O types among the peptic ulcer patients.

EXAMPLE 9.13.5—The preceding χ^2 tests may be summarized as follows:

Comparison	<i>d.f.</i>	χ^2
O, A, B types in gastric cancer (g) and controls (c)	2	5.64
A, B types in peptic ulcer and combined (g, c)	1	0.68
O, A and B types in peptic ulcer and combined (g, c)	1	34.29
Total	4	40.61

The total χ^2 , 40.61, is close to the original χ^2 , 40.54, because we have broken down the original 4 *d.f.* into a series of independent operations that account for all 4 *d.f.* The difference between 40.61 and 40.54, however, is not just a rounding error: the two quantities differ a little algebraically.

9.14—Sets of 2×2 tables. Sometimes the task is to combine the evidence from a number of 2×2 tables. The same two treatments or types of subject may have been compared in different studies, and it is desired to summarize the combined data. Alternatively, the results of a single investigation are often subclassified by the levels of a factor or variable that is thought to influence the results. The data in table 9.14.1, made available by Dr. Martha Rogers (in 9), are of this type.

The data form part of a study of the possible relationship between complications of pregnancy of mothers and behavior problems in children. The comparison is between mothers of children in Baltimore schools who had been referred by their teachers as behavior problems and mothers of control children not so referred. For each mother it was recorded whether

TABLE 9.14.1
A SET OF THREE 2×2 TABLES: NUMBERS OF MOTHERS WITH PREVIOUS INFANT LOSSES

Birth Order	Type of Children	No. of Mothers with:		Total	% Loss	χ^2 (1 <i>d.f.</i>)
		Losses	No Losses			
2	Problems	20	82	102 = n_{11}	19.6 = \hat{p}_{11}	
	Controls	10	54	64 = n_{12}	15.6 = \hat{p}_{12}	
3-4	Total	30	136	166 = n_1	18.1 = \hat{p}_1	0.42
	Problems	26	41	67 = n_{21}	38.8 = \hat{p}_{21}	
	Controls	16	30	46 = n_{22}	34.8 = \hat{p}_{22}	0.19
	Total	42	71	113 = n_2	37.2 = \hat{p}_2	
5+	Problems	27	22	49 = n_{31}	55.1 = \hat{p}_{31}	
	Controls	14	23	37 = n_{32}	37.8 = \hat{p}_{32}	
	Total	41	45	86 = n_3	47.7 = \hat{p}_3	2.52

he had suffered any infant losses (e.g., stillbirths) prior to the birth of the child. Since these loss rates increase with the birth order of the child, as table 9.14.1 shows, and since the two samples might not be comparable in the distributions of birth orders, the data were examined separately for three birth-order classes. This is a common type of precaution.

Each of the three 2×2 tables is first inspected separately. None of the χ^2 values in a single table, shown at the right, approaches the 5% significance level. Note, however, that in all three tables the percentage of mothers with previous losses is higher in the problem children than in the controls. We seek a test sensitive in detecting a population difference that is consistently in one direction, although it may not show up clearly in the individual tables.

A simple method is to compute $\hat{\chi}$ (the square root of χ^2) in each table. Give any χ_i the same sign as the difference $d_i = \hat{p}_{i1} - \hat{p}_{i2}$, and add the $\hat{\chi}_i$ values. From table 9.14.1,

$$\chi_1 + \chi_2 + \chi_3 = +0.650 + 0.436 + 1.587 = +2.673,$$

each χ_i being + because all the differences are +

Under H_0 , any χ_i is a standard normal deviate: hence, the sum of the 3 χ 's is a normal deviate with $S.D. = \sqrt{3}$. The test criterion is $\Sigma \chi_i / \sqrt{g}$, where g is the number of tables. In this case we have $2.673 / \sqrt{3} = 1.54$. In the normal table, the two-tailed P value is just above 0.10. For this test the χ 's should not be corrected for continuity.

This test is satisfactory if (i) the n_i do not vary from table to table by more than a ratio of 2 to 1, and (ii) the \hat{p}_i are in the range 20% to 75%. If the n_i vary greatly, this test gives too much weight to the small tables, which have relatively poor power to reveal a falsity in the $N.H.$ If the p 's in some tables are close to zero or 100%, while others are around 50%, the population differences δ_i are likely to be related to the level of the p_{ij} . Suppose that we are comparing the proportions of cases in which body injury is suffered in auto accidents by seat-belt wearers and non-wearers. The accidents have been classified by severity of impact into mild, moderate, severe, extreme, giving four 2×2 tables. Under the mild impacts, both p_{11} and p_{12} may be small and δ_1 also small, since injury rarely occurs with mild impact. Under extreme impact, p_{41} and p_{42} may both be close to 100%, making δ_4 also small. The large δ 's may occur in the two middle tables where the p 's are nearer 50%.

In applications of this type, two mathematical models have been used to describe how δ_i may be expected to change as p_{i2} changes. One model supposes that the difference between the two populations is constant on a *logit* scale. The logit of a proportion p is $\log_e(p/q)$. A constant difference on the logit scale means that $\log_e(p_{11}/q_{11}) - \log_e(p_{12}/q_{12})$ is constant as p_{12} varies. The second model postulates that the difference is constant on a *normal deviate* (Z) scale. The value of Z corresponding to any proportion p is such that the area of a standard normal curve to the

left of Z is p . For instance, $Z = 0$ for $p = 0.5$, $Z = 1.282$ for $p = 0.9$, $Z = -1.282$ for $p = 0.1$.

To illustrate the meaning of a constant difference on these transformed scales, table 9.14.2 shows the size of difference on the original percentage scale that corresponds to a constant difference on (a) the logit scale (b) the normal deviate scale. The size of the difference was chosen to equal 20% at $p_2 = 50\%$. Note that (i) the differences diminish towards both ends of the p scale as in the seat belts example, (ii) the two transformations do not differ greatly.

TABLE 9.14.2
SIZE OF DIFFERENCE $\delta = p_1 - p_2$ FOR A RANGE OF VALUES OF p_2

$p_2\%$	1	5	10	30	50	70	90	95	99
Constant logit	2.6	8.1	12.4	20.0	20.0	15.3	6.4	3.5	0.8
Constant Z	1.3	6.0	10.6	20.0	20.0	14.5	5.5	2.8	0.6

A test that gives appropriate weight to tables with large n_i and is sensitive if differences are constant on a logit or a Z scale was developed by Cochran (9). If \hat{p}_i is the combined percentage in the i th table, and

$$w_i = n_{i1}n_{i2}/(n_{i1} + n_{i2}) \quad : \quad d_i = \hat{p}_{i1} - \hat{p}_{i2},$$

we compute

$$\Sigma w_i d_i / \sqrt{\Sigma w_i \hat{p}_i \hat{q}_i}$$

and refer to the normal table. For the data in table 9.14.1 the computations are as follows (with the d_i in proportions to keep the numbers smaller).

Birth Order	n_i	d_i	$n_i d_i$	\hat{p}_i	$\hat{p}_i \hat{q}_i$	$w_i \hat{p}_i \hat{q}_i$
2	39.3	+0.040	+1.57	0.181	0.1482	5.824
3-4	27.3	+0.040	+1.09	0.372	0.2336	6.377
5+	21.1	+0.173	+3.65	0.477	0.2494	5.262
Sum			+6.31			17.463

The test criterion is $6.31/\sqrt{17.463} = 1.51$. This agrees closely with the value 1.54 found by the $\Sigma\chi$ test, for which these tables are quite suitable.

There is another way of computing this test. In the i th table, let O_i be the observed number of Problems losses and E_i the expected number under H_0 . For birth order 2 (table 9.14.1), $O_i = 20$, $E_i = (30)(102)/166$

TABLE 9.14.3
THE MANTEL-HAENSZEL TEST FOR THE INFANT LOSS DATA IN TABLE 9.14.1

Birth Order	O_i	E_i	$n_{i1}n_{i2}c_{i1}c_{i2}/n_i^2(n_i - 1)$
2	20	18.43	5.858
3-4	26	24.90	6.426
5+	27	23.36	5.321
sum	73	66.69	17.605
$Z = (73 - 66.69 - \frac{1}{2})/\sqrt{17.605} = 1.38$			

= 18.43. Then $(O_1 - E_1) = +1.57$, which is the same as w_1d_1 . This result may be shown by algebra to hold in any 2×2 table. The criterion can therefore be written

$$\Sigma(O_i - E_i)/\sqrt{\Sigma w_i \hat{p}_i \hat{q}_i}$$

This form of the test has been presented by Mantel and Haenszel (28, 29), with two refinements that are worthwhile when the n 's are small. First, the variance of $w_i d_i$ or $(O_i - E_i)$ on H_0 is not $w_i \hat{p}_i \hat{q}_i$ but the slightly larger quantity $n_{i1}n_{i2}\hat{p}_i\hat{q}_i/(n_{i1} + n_{i2} - 1)$. If the margins of the 2×2 table are n_{i1} , n_{i2} , c_{i1} , and c_{i2} , this variance can be computed as

$$n_{i1}n_{i2}c_{i1}c_{i2}/n_i^2(n_i - 1), \quad (n_i = n_{i1} + n_{i2}),$$

a form that is convenient in small tables.

Secondly, a correction for continuity can be applied by subtracting $1/2$ from the absolute value of $\Sigma(O_i - E_i)$. This version of the test is shown in table 9.14.3. The correction for continuity makes a noticeable difference even with samples of this size.

The analysis of proportions is discussed further in sections 16.8-16.12.

REFERENCES

1. E. W. LINDSTROM. *Cornell Agric. Exp. Sta.*, Memoir 13 (1918).
2. A. W. F. EDWARDS. *Ann. Hum. Gen.*, 24:309 (1960).
3. J. H. EDWARDS. *Ann. Hum. Gen.*, 25:89 (1961).
4. C. W. LEGGATT. *Comptes rendus de l'association internationale d'essais de semences*, 5:27 (1935).
5. D. J. CAFFEY and C. E. SMITH. Bureau of Entomology and Plant Quarantine, USDA (Baton Rouge) (1934).
6. W. G. COCHRAN. *Ann. Math. Statist.*, 23:315 (1952).
7. I. M. CHAKRAVARTI and C. R. RAO. *Sankhyā*, 21:315 (1959).
8. W. G. COCHRAN. *J. R. Statist. Soc. Suppl.*, 3:49 (1936).
9. W. G. COCHRAN. *Biometrics*, 10:417 (1954).
10. G. W. SNEDECOR and M. R. IRWIN. *Iowa State Coll. J. Sci.*, 8: 75 (1933).
11. R. C. LEWONTIN and J. FELSENSTEIN. *Biom.*, 21:19 (1965).
12. J. B. S. HALDANE. *Biometrika*, 33:234 (1943-46).
13. J. O. IRWIN and E. A. CHEESEMAN. *J. R. Statist. Soc. Suppl.* 6:174 (1939).
14. L. C. BURNETT. M.S. Thesis. Iowa State College (1906).
15. G. C. DECKER and F. ANDRE. *Iowa State J. Sci.*, 10:403 (1936).

16. A. W. KIMBALL. *Biometrics*, 10:452 (1952).
17. D. J. BARTHOLOMEW. *Biometrika*, 46:328 (1959).
18. F. YATES. *Biometrika*, 35:176 (1948).
19. P. ARMITAGE. *Biometrics*, 11:375 (1955).
20. K. PEARSON. *Biometrika*, 5:105 (1905-06).
21. I. D. J. BROSS. *Biometrics*, 14:18 (1958).
22. J. IPSEN. *Biometrics*, 12:465 (1955).
23. R. A. FISHER. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh (1941).
24. R. P. ABELSON and J. W. TUKEY. *Proc. Soc. Statist. Sect. Amer. Statist. Ass.* (1959).
25. R. A. FISHER. *Ann. Sci.*, 1:117 (1936).
26. N. V. STRAND and R. J. JESSEN. *Iowa Agr. Exp. Stat. Res. Bul.* 315 (1943).
27. P. H. LESLIE. *Biometrics*, 7: 283 (1951).
28. N. MANTEL and W. HAENSZEL. *J. Nat. Cancer Inst.*, 22:719 (1959).
29. N. MANTEL. *J. Amer. Statist. Ass.*, 58:690 (1963).

One-way classifications. Analysis of variance

10.1—Extension from two samples to many. Statistical methods for two independent samples were presented in chapter 4, but the needs of the investigator are seldom confined to the comparison of two samples only. For attribute data, the extension to more than two samples was made in the preceding chapter. We are now ready to do the same for measurement data.

First, recall the analysis used in the comparison of two samples. In the numerical example (section 4.9, p. 102), the comb weights of two samples of 11 chicks were compared, one sample having received sex hormone A, the other sex hormone C. Briefly, the principal steps in the analysis were as follows: (i) the mean comb weights \bar{X}_1, \bar{X}_2 were computed, (ii) the within-sample sum of squares of deviations Σx^2 , with 10 *df.*, was found for each sample, (iii) a pooled estimate s^2 of the within-sample variance was obtained by adding the two values of Σx^2 and dividing by the sum of the *df.*, 20, (iv) the standard error of the mean difference, $\bar{X}_1 - \bar{X}_2$, was calculated as $\sqrt{(2s^2/n)}$, where $n = 11$ is the size of each sample, (v) finally, a test of the null hypothesis $\mu_1 = \mu_2$ and confidence limits for $\mu_1 - \mu_2$ were given by the result that the quantity

$$\{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)\} / \sqrt{(2s^2/n)}$$

follows the *t*-distribution with 20 *df.*

In the next section we apply this method to an experiment with four treatments, i.e., four independent samples.

10.2—An experiment with four samples. During cooking, doughnuts absorb fat in various amounts. Lowe (1) wished to learn if the amount absorbed depends on the type of fat used. For each of four fats, six batches of doughnuts were prepared, a batch consisting of 24 doughnuts. The data in table 10.2.1 are the grams of fat absorbed per batch, coded by deducting 100 grams to give simpler figures. Data of this kind are called a *single* or *one-way* classification, each fat representing one class.

Before beginning the analysis, note that the totals for the four fats differ substantially, from 372 for fat 4 to 510 for fat 2. Indeed, there is a

In a second sample of 500 adults, handled in the same manner except that they were exposed at $-9^{\circ}\text{C}.$, the numbers dead in groups of 100 were 38, 30, 30, 40, 27. The χ^2 value of 5.79 again verifies the technique, showing only sampling variation from the estimated mortality of 33%.

The gratifying uniformity in the results leads one to place some confidence in the surprising finding that the death rates at $-8^{\circ}\text{C}.$ and $-9^{\circ}\text{C}.$ were markedly different. The total numbers dead in the two samples of 500 were 88 and 165. The result, $\chi^2 = 31.37$ with $d.f. = 1$, P less than 0.0002, provides convincing evidence that a rise in mortality with the lowering of temperature from $-8^{\circ}\text{C}.$ to $-9^{\circ}\text{C}.$ is a characteristic of the population, not merely an accident of sampling.

The ease of applying a test of experimental technique makes its use almost a routine procedure except in highly standardized processes. It is necessary merely to collect the data in several small groups, chosen with regard to the types of experimental variation thought likely to be present, instead of in one mass. The additional information may modify conclusions and subsequent procedures profoundly.

In this example the sum of the three values of χ^2 is $4.22 + 5.79 + 31.37 = 41.38$, with 9 $d.f.$ If the initial χ^2 is calculated from the 2×10 contingency table formed by the complete data, its value is also found to be 41.38, with 9 $d.f.$ This agreement between the two values is a fluke, which does not hold generally in $2 \times C$ tables. For $2 \times C$ and $R \times C$ tables, a method of computing the component parts so that they add to the initial total χ^2 is available (16). In these data this method amounts to using the same denominator $\bar{p}\bar{q} = (0.253)(0.747)$, calculated from the *total* mortality, in finding all χ^2 values. Instead, for the 4 $d.f.$ χ^2 at $-8^{\circ}\text{C}.$ we used $\bar{p}\bar{q} = (0.176)(0.824)$, appropriate to that part of the data, and at $-9^{\circ}\text{C}.$ we used $\bar{p}\bar{q} = (0.330)(0.670)$. The additive χ^2 values give $3.24 + 6.77 + 31.37 = 41.38$. However, when it has been shown that the mortality differs at $-8^{\circ}\text{C}.$ and $-9^{\circ}\text{C}.$, use of a pooled \bar{p} for the individual homogeneity tests at $-8^{\circ}\text{C}.$ and $-9^{\circ}\text{C}.$ is invalid. The non-additive method is recommended, except in a quick preliminary look at the data.

9.10—Ordered classifications. In the leprosy example of section 9.7, the classes (marked improvement, moderate improvement, slight improvement, stationary, worse) are an example of an *ordered classification*. Such classifications are common in the study of human behavior and preferences, and more generally whenever different degrees of some phenomenon can be recognized. The problem of utilizing the knowledge that we possess about this ordering has attracted considerable attention in recent years.

With a single classification of Poisson variables, the ordering might lead us to expect that if the null hypothesis $\mu_i = \mu$ does not hold, an alternative $\mu_1 \leq \mu_2 \leq \mu_3 \leq \dots$ should hold, where the subscripts represent the order. For instance, if working conditions in a factory have been classified as Excellent, Good, Fair, we might expect that if the number of defective articles per worker varies with working conditions, the order should

be $\mu_1 \leq \mu_2 \leq \mu_3$. Similarly, with ordered columns in a $2 \times C$ contingency table, the alternative $p_1 \leq p_2 \leq p_3 \leq \dots$ might be expected. χ^2 tests designed to detect this type of alternative have been developed by Bartholomew (17). The computations are quite simple.

Another approach, used by numerous workers (9), (18), (19), is to attach a score to each class so that an ordered scale is created. To illustrate from the leprosy example, we assigned scores of 3, 2, 1, respectively, to the Marked, Moderate, and Slight Improvement classes, 0 to the Stationary class, and -1 to the Worse class. These scores are based on the judgment that the five classes constructed by the expert represent equal gradations on a continuous scale. We considered giving a score of $+4$ to the Marked Improvement class and -2 to the Worse class, since the expert seemed to examine a patient at greater length before assigning him to one of these extreme classes, but rejected this since our impression may have been erroneous.

Having assigned the scores we may think of the leprosy data as consisting of two independent samples of 144 and 52 patients, respectively. (See table 9.10.1.) For each patient we have a discrete measure X of his change in health, where X takes only the values 3, 2, 1, 0, -1 . We can estimate the average change in health for each sample, with its standard error, and can test the null hypothesis that this average change is the same in the two populations. For this test we use the ordinary two-sample t -test as applied to grouped data. The calculations appear in table 9.10.1. On the X scale the average change in health is $+1.269$ for patients with much infiltration and $+0.819$ for those with little infiltration. The difference, \bar{D} , is 0.450, with standard error ± 0.172 (194 *d.f.*), computed in the usual way. The value of t is $0.450/0.172 = 2.616$, with $P < 0.01$. Contrary to the initial χ^2 test, this test reveals a significantly greater amount of progress for the patients with much infiltration.

The assignment of scores is appropriate when (i) the phenomenon in question is one that could be measured on a continuous scale if the instruments of measurement were good enough, and (ii) the ordered classification can be regarded as a kind of grouping of this continuous scale, or as an attempt to approximate the continuous scale by a cruder scale that is the best we can do in the present state of knowledge. The process is similar to that which occurs in many surveys. The householder is shown five specific income classes and asked to indicate the class within which his income falls, without naming his actual income. Some householders name an incorrect class, just as an expert makes some mistakes in classification when this is difficult.

The advantage in assigning scores is that the more flexible and powerful methods of analysis that have been developed for continuous variables become available. One can begin to think of the sizes of the average differences between different groups in a study, and compare the difference between groups A and B with that between groups E and F . Regressions of the group means \bar{X} on a further variable Z can be worked

TABLE 9.10.1
ANALYSIS OF THE LEPROSY DATA BY ASSIGNED SCORES
(Data with assigned scores)

Change in Health	Infiltration	
	Little	Much
X	<i>No. of patients</i>	
	f	f
3	11	7
2	27	15
1	42	16
0	53	13
-1	11	1
Total: Σf	144	52

(Computations)

	<i>Little</i>	<i>Much</i>
ΣfX	118	66
$\bar{X} = \Sigma fX / \Sigma f$	0.819	1.269
ΣfX^2	260	140
$(\Sigma fX)^2 / \Sigma f$	96.7	83.8

Σfx^2	163.3	56.2
$d.f.$	143	51
s^2	1.142	1.102
Pooled s^2	1.131	

$$s_D^2 = (1.131) \left(\frac{1}{144} + \frac{1}{52} \right) = 0.0296$$

$$s_D = 0.172$$

$$t = \frac{\bar{D}}{s_D} = \frac{1.269 - 0.819}{0.172} = 2.616$$

$$d.f. = 194, \quad P < 0.01$$

out. The relative variability of different groups can be examined by computing s for each group.

This approach assumes that the standard methods of analysis of continuous variables, like the t -test, can be used with an X variable that is discrete and takes only a few values. As noted in section 5.8 on scales with limited values, the standard methods appear to work well enough for practical use. However, heterogeneity of variance and correlation between s^2 and \bar{X} are more frequently encountered because of the discrete scale. If most of the patients in a group show marked improvement, most of their X 's will be 3, and s^2 will be small. Pooling of variances should not be undertaken without examining the individual s^2 . In the leprosy example the two s^2 were 1.142 and 1.102 (table 9.10.1), and this difficulty was not present.

The chief objection to the assignment of scores is that the method is more or less arbitrary. Two investigators may assign different scores to the same set of data. In our experience, however, moderate differences between two scoring systems seldom produce marked differences in the conclusions drawn from the analysis. In the leprosy example, the alternative scores 4, 2, 1, 0, -2 give $t = 2.549$ as against $t = 2.616$ in the analysis in table 9.10.1. Some classifications present particular difficulty. If the degrees of injury to persons in accidents are recorded as slight, moderate, severe, disabling, and fatal, there seems no entirely satisfactory way of placing the last two classes on the same scale as the first three.

Several alternative principles have been used to construct scores. In studies of different populations of school children, K. Pearson (20) assumed that the underlying continuous variate was normally distributed in a standard population of school children. If the classes are regarded as a grouping of this normal distribution, the class boundaries for the normal variate are easily found. The score assigned to a class is the mean of the normal variate within the class. A related approach due to Bross (21) also uses a standard population but does not assume normality. The score (*ridit*) given to a class is the relative frequency up to the midpoint of that class in the standard population. When the experimental treatments are different doses of a toxic or protective agent in biological assay, Ipsen (22) shows how to assign scores so that the resulting variate has a linear regression on some chosen function of the dose, the ratio of the variance due to regression to the total variance being maximized. Fisher (23) assigns scores so as to maximize the F -ratio of treatments to experimental error as defined in section 10.5. The *maximin* method of Abelson and Tukey (24), maximizes the square of the correlation coefficient r^2 between the assigned scores and the set of true scores, consistent with the investigator's knowledge about the ordering of the classes, that gives a minimum correlation with the assigned scores. This approach, like Bartholomew's, avoids any arbitrary assumptions about the nature of the true scale.

EXAMPLE 9 10 1—In the leprosy data, verify the value of $t = 2.549$ quoted for the scoring 4, 2, 1, 0, -2

9.11—Test for a linear trend in proportions. When interest is centered on the proportions p_i in a $2 \times C$ contingency table, there is another way of viewing the data. Table 9.11.1 shows the leprosy data with the assigned scores X_i , but in this case the variable that we analyze is p_i , the proportion of patients with much infiltration. The contention now is that if these patients have fared better than patients with little infiltration, the values of p_i should increase as we move from the Worse class ($X = -1$) towards the Marked Improvement class ($X = 3$).

If this is so, the regression coefficient of p_i on X_i should be a good test criterion. On the null hypothesis (no relation between p_i and X_i) each p_i is distributed about the same mean, estimated by \bar{p} , with variance $\bar{p}\bar{q}/n_i$. The regression coefficient b is calculated as usual, except that each p_i must be weighted by the reciprocal of the sample size n_i on which it is

TABLE 9 11 1
TESTING A LINEAR REGRESSION OF p_i ON THE SCORE (LEPROSY DATA)

Degree of Infiltration	Improvement			Stationary	Worse	Total
	Marked	Moderate	Slight			
Little	11	27	42	53	11	144
Much (a_i)	7	15	16	13	1	52
Total (n_i)	18	42	58	66	12	196 (N)
$p_i = a_i/n_i$	0.3889	0.3571	0.2759	0.1970	0.0833	0.2653 (p)
Score X_i	3	2	1	0	-1	

based. The numerator and denominator of b are computed as follows:

$$\begin{aligned}
 \text{Num.} &= \sum n_i(p_i - \bar{p})(X_i - \bar{X}) \\
 &= \sum n_i p_i X_i - (\sum n_i p_i)(\sum n_i X_i)/\sum n_i \\
 &= \sum a_i X_i - (\sum a_i)(\sum n_i X_i)/N \\
 &= 66 - (52)(184)/196 = 66 - 48.82 = 17.18
 \end{aligned}$$

$$\begin{aligned}
 \text{Den.} &= \sum n_i X_i^2 - (\sum n_i X_i)^2/N \\
 &= 400 - (184)^2/196 = 400 - 172.8 = 227.2
 \end{aligned}$$

This gives $b = 17.18/227.2 = 0.0756$. Its standard error is

$$s_b = \sqrt{(\bar{p}\bar{q}/\text{Den.})} = \sqrt{\{(0.2653)(0.7347)/(227.2)\}} = 0.0293$$

The normal deviate for testing the null hypothesis $\beta = 0$ is

$$Z = b/s_b = 0.0756/0.0293 = 2.580. \quad P = 0.0098.$$

Although it is not obvious at first sight, Yates (18) showed that this regression test is essentially the same as the t -test in section 9.10 of the difference between the mean scores in the Little and Much infiltration classes. In this example the regression test gave $Z = 2.580$ while the t -test gave $t = 2.616$ (194 d.f.). The difference in results arises because the two approaches use slightly different large-sample approximations to the exact distributions of Z and t with these discrete data.

EXAMPLE 9 11 1 Armitage (19) quotes the following data by Holmes and Williams for the relation in children between size of tonsils and the proportion of children who are carriers of *streptococcus pyogenes* in the nose

Types of Children	X = Score Given to Size of Tonsils			Total Children
	0	1	2	
Carriers (a_i)	19	29	24	72 (A)
Non-carriers	497	560	269	1326
Total (n_i)	516	589	293	1398 (N)
Carrier-rate (p_i)	0.0368	0.0492	0.0819	0.051502 (\bar{p})

Calculate. (i) the normal deviate Z for testing the linear regression of the proportion of carriers on size of tonsils, (ii) the value of t for comparing the difference between the mean size of tonsils in carriers and non-carriers. Ans. (i) $Z = 2.681$, (ii) $t = 2.686$, with 1396 $d.f.$

EXAMPLE 9.11.2—When the regression of p_i on X_i is used as a test criterion, it is of interest to examine whether the regression is linear. Armitage (19) shows that this can be done by first computing $\chi^2 = \sum n_i(p_i - \bar{p})^2 / \bar{p}\bar{q} = \{\sum a_i p_i - A^2/N\} / \bar{p}\bar{q}$. This χ^2 , with $(C - 1)$ $d.f.$, measures the total variation among the C values of p_i . The χ^2 for linear regression, with 1 $d.f.$, is found by squaring Z , since the square of a normal deviate has a χ^2 distribution with 1 $d.f.$ The difference, $\chi_{(C-1)}^2 - \chi_1^2$, is a χ^2 with $(C - 2)$ $d.f.$ for testing the deviations of the p_i from their linear regression on the X_i . Compute this χ^2 for the data in example 9.11.1. Ans. The total χ^2 is 7.85 with 2 $d.f.$, while Z^2 is 7.19 with 1 $d.f.$ Thus the χ^2 for the deviations is 0.66 with 1 $d.f.$, in agreement with the hypothesis of linearity.

9.12—Heterogeneity χ^2 in testing Mendelian ratios. It is often advisable to collect data in several small samples rather than in a single large one. An example is furnished by some experiments on chlorophyll inheritance in maize (1), reported in table 9.12.1. The series consisted of 11 samples of progenies of heterozygous green plants, self-fertilized, segregating into dominant green plants and recessive yellow plants. The hypothetical ratio is 3 green to 1 yellow. We shall study the proportion of yellow—theoretically 1/4.

TABLE 9.12.1
NUMBER OF YELLOW SEEDLINGS IN 11 SAMPLES OF MAIZE

No. in Sample	No. Yellow	Proportion Yellow
n_i	a_i	$p_i = a_i/n_i$
122	24	0.1967
149	39	0.2617
86	18	0.2093
55	13	0.2364
71	17	0.2394
179	38	0.2123
150	30	0.2000
36	9	0.2500
91	21	0.2308
53	14	0.2642
111	26	0.2342
$N = 1103$	$A = 249$	$\bar{p} = 0.22575$

Heterogeneity χ^2 (10 $d.f.$)

$$\chi^2 = (\sum a_i p_i - A\bar{p}) / \bar{p}\bar{q} = (0.5779) / (0.2258)(0.7742) = 3.31$$

Pooled χ^2 (1 $d.f.$)

$$\chi_c^2 = (|A - N\bar{p}| - \frac{1}{2})^2 / N\bar{p}\bar{q} \\ = (|249 - 275.75| - \frac{1}{2})^2 / (1103)(0.25)(0.75) = 3.33$$

The data may fail to satisfy the simple Mendelian hypothesis in two ways. First, there may be real differences among the p_i (proportion of yellow) in different samples. This finding points to some additional source

of variability that must be explained before the data can be used as a crucial test of the Mendelian ratio. Second, the p_i may agree with one another (apart from sampling errors) but their overall proportion \bar{p} may disagree with the Mendelian proportion p . The reason may be linkage or crossing-over, or differential robustness in the dominant and recessive plants.

The first point is examined by applying to the p_i the variance test for homogeneity of the binomial distribution (section 9.8). The value of χ^2 shown under table 9.12.1, is 3.31, with 10 *d.f.*, P about 0.97. The test gives no reason to suspect real differences among the p_i . We therefore pool the samples and compare the overall ratio, $\bar{p} = 0.22575$, with the hypothetical $p = 0.25$, by the χ^2 test for a binomial proportion (section 8.8). We find χ^2 (corrected for continuity) = 3.33, P about 0.07. There is a hint of a deficiency of the recessive yellows.

In showing the relation between these two tests, the following algebraic identity is of interest:

$$\frac{\sum_i^C n_i(p_i - p)^2}{pq} = \frac{(\sum_i^C n_i)(\bar{p} - p)^2}{pq} + \frac{\sum_i^C n_i(p_i - \bar{p})^2}{pq} \quad (9.12.1)$$

The quantity $n_i(p_i - p)^2/pq$ measures the discrepancy between the observed p_i in the i th sample and the theoretical value p . If the null hypothesis is true, this quantity is distributed as χ^2 with 1 *d.f.* and the sum of these quantities over the C samples (left side of equation 9.12.1) is distributed as χ^2 with C *d.f.* The first term on the right of (9.12.1) compares the pooled ratio \bar{p} with p , and is distributed as χ^2 with 1 *d.f.* The second term on the right measures the deviations of the p_i from their own pooled mean \bar{p} , and is distributed as χ^2 with $(C - 1)$ *d.f.* To sum up, the total χ^2 on the left, with C *d.f.*, splits into a χ^2 with 1 *d.f.* which compares the pooled sample \bar{p} and the theoretical p , and a heterogeneity χ^2 , with $(C - 1)$ *d.f.*, which compares the p_i among themselves. These χ^2 distributions are of course followed only approximately unless the n_i are large.

In practice, this additive feature is less useful. Unless the pooled sample is large, a correction for continuity in the 1 *d.f.* for the pooled χ^2 is advisable. This destroys the additivity. Secondly, the expression for the heterogeneity χ^2 assumes that the theoretical ratio p applies in these data. If there is doubt on this point, the heterogeneity χ^2 should be calculated, as in table 9.12.1, with $\bar{p}\bar{q}$ in the denominator instead of pq . In this form the heterogeneity χ^2 involves no assumption that $\bar{p} = p$ (apart from sampling errors).

EXAMPLE 9.12.1—From a population expected to segregate 1:1, four samples with the following ratios were drawn, 47:33, 40:26, 30:42, 24:34. Note the discrepancies among the sample ratios. Although the pooled χ^2 does not indicate any unusual departure from the theoretical ratio, you will find a large heterogeneity χ^2 equal to 9.01, $P = 0.03$, for which some explanation should be sought.

EXAMPLE 9.12.2—Fisher (25) applied χ^2 tests to the experiments conducted by Mendel in 1863 to test different aspects of his theory, as follows:

Experiment	χ^2	d.f.
Trifactorial	8.94	17
Bifactorial	2.81	8
Gametic ratios	3.67	15
Repeated 2:1 test	0.13	1

Show that in random sampling the probability of obtaining a total χ^2 lower than that observed is less than 0.005 (use the χ^2 table). More accurately, the probability is less than 1 in 2000. Thus, the agreement of the results with Mendel's laws looks too good to be true. Fisher gives an interesting discussion of possible reasons.

9.13—The $R \times C$ table. If each member of a sample is classified by one characteristic into R classes, and by a second characteristic into C classes, the data may be presented in a table with R rows and C columns. The entry in any of the RC cells is the number of members of the sample falling into that cell. Strand and Jessen (26) classified a random sample of farms in Audubon County, Iowa, into three classes (Owned, Rented, Mixed), according to the tenure status and into three classes (I, II, III), according to the level of the soil fertility (table 9.13.1).

TABLE 9.13.1
NUMBERS OF FARMS ON THREE SOIL FERTILITY GROUPS IN AUDUBON COUNTY, IOWA,
CLASSIFIED ACCORDING TO TENURE

Soil		Owned	Rented	Mixed	Total
I	f	36	67	49	152
	F	36.75	62.92	52.33	
	$f - F$	-0.75	4.08	-3.33	
II	f	31	60	49	140
	F	33.85	57.95	48.20	
	$f - F$	-2.85	2.05	0.80	
III	f	58	87	80	225
	F	54.40	93.13	77.47	
	$f - F$	3.60	-6.13	2.53	
Total		125	214	178	517

$$\chi^2 = \sum \frac{(f - F)^2}{F} = \frac{(-0.75)^2}{36.75} + \dots + \frac{(2.53)^2}{77.47} = 1.54, \text{ d.f.} = (R - 1)(C - 1) = 4$$

Before drawing conclusions about the border totals for tenure status, this question is asked: Are the relative numbers of Owned, Rented, and Mixed farms in this county the same at the three levels of soil fertility?

This question might alternatively be phrased: Is the distribution of the soil fertility levels the same for Owned, Rented, and Mixed farms? (If a little reflection does not make it clear that these two questions are equivalent, see example 9.13.1.) Sometimes the question is put more succinctly as: Is tenure status *independent* of fertility level?

The χ^2 test for the $2 \times C$ table extends naturally to this situation. As before,

$$\chi^2 = \Sigma(f - F)^2/F,$$

where f is the observed frequency in any cell and F the frequency expected if the null hypothesis of independence holds.

As before, the expected frequency for any cell is computed from the border totals in the corresponding row and column:

$$\begin{aligned} F &= \frac{(\text{row total})(\text{column total})}{n} \\ &= \frac{\text{row total}}{n} (\text{column total}) \end{aligned}$$

Examples: For the first row,

$$\begin{aligned} \frac{\text{row total}}{n} &= \frac{152}{517} = 0.29400 \\ F_1 &= (0.29400)(125) = 36.75 \\ F_2 &= (0.29400)(214) = 62.92 \\ F_3 &= (0.29400)(178) = 52.33 \end{aligned}$$

This procedure makes the computation easy with a calculating machine. For verification, notice that (i) the sum of the F in any row or column is equal to the observed total, and consequently (ii) the sum of the deviations in each row and in each column is zero.

The facts just stated dictate the number of degrees of freedom. One is free to put $R - 1$ expected frequencies in a column, but the remaining cell is then fixed as the column total minus the sum of the $R - 1$ values of F . Similarly, when we have inserted expected frequencies in this way in $(C - 1)$ columns, the expected frequencies in the last column are fixed. Therefore, $d.f. = (R - 1)(C - 1)$.

The calculation of χ^2 is given in the table. Since $P > 0.8$, the null hypothesis is not rejected. If you do not need to examine the contribution of the individual cells of χ^2 , up to half the time in computation can be saved by a shortcut devised by P. H. Leslie (27). This is especially useful if many tables are to be calculated.

When χ^2 is significant, the next step is to study the nature of the departure from independence in more detail. Examination of the cells in which the contribution to χ^2 is greatest, taking note of the signs of the deviations $(f - F)$, furnishes clues, but these are hard to interpret because the deviations in different cells are correlated. Computation of the per-

centage distribution of the row classification within each column, followed by a scrutiny of the changes from column to column, may be more informative. Further χ^2 tests may help. For instance, if the percentage distribution of the row classification appears the same in two columns, a χ^2 test for these two columns may confirm this. The two columns can then be combined for comparison with other columns. Examples 9.13.2, 3, 4, 5 illustrate this approach.

EXAMPLE 9.13.1—Show that if the expected distribution of the column classification is the same in every row, then the expected distribution of the row classification is the same in every column. For the i th row, let $F_{i1}, F_{i2}, \dots, F_{ic}$ be the expected numbers in the respective columns. Let $F_{i2} = a_2 F_{i1}, F_{i3} = a_3 F_{i1}, \dots, F_{ic} = a_c F_{i1}$. Then the numbers a_2, a_3, \dots, a_c must be the same in every row, since the expected distribution of the column classification is the same in every row. Now the expected row distribution in the first column is $F_{11}, F_{21}, \dots, F_{R1}$. In the second column it is $F_{12} = a_2 F_{11}, F_{22} = a_2 F_{21}, \dots, F_{R2} = a_2 F_{R1}$. Since a_2 is a constant multiplier, this is the same distribution as in the first column, and similarly for any other column.

EXAMPLE 9.13.2—In a study of the relation between blood type and disease, large samples of patients with peptic ulcer, patients with gastric cancer, and control persons free from these diseases were classified as to blood type (O, A, B, AB). In this example, the relatively small numbers of AB patients were omitted for simplicity. The observed numbers are as follows

Blood Type	Peptic Ulcer	Gastric Cancer	Controls	Totals
<i>O</i>	983	383	2892	4528
<i>A</i>	679	416	2625	3720
<i>B</i>	134	84	570	788
Totals	1796	883	6087	8766

Compute χ^2 to test the null hypothesis that the distribution of blood types is the same for the three samples. Ans. $\chi^2 = 40.54$, 4 d.f. P very small.

EXAMPLE 9.13.3—To examine this question further, compute the percentage distribution of blood types for each sample, as shown below.

Blood Type	Peptic Ulcer	Gastric Cancer	Controls
<i>O</i>	54.7	43.4	47.5
<i>A</i>	37.8	47.1	43.1
<i>B</i>	7.5	9.5	9.4
Totals	100.0	100.0	100.0

This suggests (i) there is little difference between the blood type distributions for gastric cancer patients and controls, (ii) peptic ulcer patients differ principally in having an excess of patients of type O . Going back to the frequencies in example 9.13.2, test the hypothesis that the blood type distribution is the same for gastric cancer patients and controls. Ans. $\chi^2 = 5.64$ (2 d.f.). P about 0.06.

EXAMPLE 9.13.4—Combine the gastric cancer and control samples. Test (i) whether the distribution of A and B types is the same in this combined sample as in the peptic ulcer sample (omit the O types). Ans. $\chi^2 = 0.68$ (1 d.f.) $P > 0.7$. (ii) Test whether proportion

of O types versus $A + B$ types is the same for the combined sample as for the gastric cancer samples. Ans. $\chi^2 = 34.29$ (1 *d.f.*). P very small. To sum up, the high value of the original 4 *d.f.* χ^2 is due primarily to an excess of O types among the peptic ulcer patients.

EXAMPLE 9.13.5—The preceding χ^2 tests may be summarized as follows:

Comparison	<i>d.f.</i>	χ^2
O, A, B types in gastric cancer (g) and controls (c)	2	5.64
A, B types in peptic ulcer and combined (g, c)	1	0.68
O, A and B types in peptic ulcer and combined (g, c)	1	34.29
Total	4	40.61

The total χ^2 , 40.61, is close to the original χ^2 , 40.54, because we have broken down the original 4 *d.f.* into a series of independent operations that account for all 4 *d.f.* The difference between 40.61 and 40.54, however, is not just a rounding error: the two quantities differ a little algebraically.

9.14—Sets of 2×2 tables. Sometimes the task is to combine the evidence from a number of 2×2 tables. The same two treatments or types of subject may have been compared in different studies, and it is desired to summarize the combined data. Alternatively, the results of a single investigation are often subclassified by the levels of a factor or variable that is thought to influence the results. The data in table 9.14.1, made available by Dr. Martha Rogers (in 9), are of this type.

The data form part of a study of the possible relationship between complications of pregnancy of mothers and behavior problems in children. The comparison is between mothers of children in Baltimore schools who had been referred by their teachers as behavior problems and mothers of control children not so referred. For each mother it was recorded whether

TABLE 9.14.1
A SET OF THREE 2×2 TABLES: NUMBERS OF MOTHERS WITH PREVIOUS INFANT LOSSES

Birth Order	Type of Children	No. of Mothers with:		Total	% Loss	χ^2 (1 <i>d.f.</i>)
		Losses	No Losses			
2	Problems	20	82	102 = n_{11}	19.6 = \hat{p}_{11}	
	Controls	10	54	64 = n_{12}	15.6 = \hat{p}_{12}	
3-4	Total	30	136	166 = n_1	18.1 = \hat{p}_1	0.42
	Problems	26	41	67 = n_{21}	38.8 = \hat{p}_{21}	
	Controls	16	30	46 = n_{22}	34.8 = \hat{p}_{22}	
	Total	42	71	113 = n_2	37.2 = \hat{p}_2	0.19
5+	Problems	27	22	49 = n_{31}	55.1 = \hat{p}_{31}	
	Controls	14	23	37 = n_{32}	37.8 = \hat{p}_{32}	
	Total	41	45	86 = n_3	47.7 = \hat{p}_3	2.52

she had suffered any infant losses (e.g., stillbirths) prior to the birth of the child. Since these loss rates increase with the birth order of the child, as table 9.14.1 shows, and since the two samples might not be comparable in the distributions of birth orders, the data were examined separately for three birth-order classes. This is a common type of precaution.

Each of the three 2×2 tables is first inspected separately. None of the χ^2 values in a single table, shown at the right, approaches the 5% significance level. Note, however, that in all three tables the percentage of mothers with previous losses is higher in the problem children than in the controls. We seek a test sensitive in detecting a population difference that is consistently in one direction, although it may not show up clearly in the individual tables.

A simple method is to compute $\bar{\chi}$ (the square root of χ^2) in each table. Give any χ_i the same sign as the difference $d_i = \hat{p}_{i1} - \hat{p}_{i2}$, and add the χ_i values. From table 9.14.1,

$$\chi_1 + \chi_2 + \chi_3 = +0.650 + 0.436 + 1.587 = +2.673,$$

each χ_i being + because all the differences are +

Under H_0 , any χ_i is a standard normal deviate: hence, the sum of the 3 χ 's is a normal deviate with $S.D. = \sqrt{3}$. The test criterion is $\Sigma \chi_i / \sqrt{g}$, where g is the number of tables. In this case we have $2.673 / \sqrt{3} = 1.54$. In the normal table, the two-tailed P value is just above 0.10. For this test the χ 's should not be corrected for continuity.

This test is satisfactory if (i) the n_i do not vary from table to table by more than a ratio of 2 to 1, and (ii) the \hat{p}_i are in the range 20% to 75%. If the n_i vary greatly, this test gives too much weight to the small tables, which have relatively poor power to reveal a falsity in the $N.H.$ If the p 's in some tables are close to zero or 100%, while others are around 50%, the population differences δ_i are likely to be related to the level of the p_{ij} . Suppose that we are comparing the proportions of cases in which body injury is suffered in auto accidents by seat-belt wearers and non-wearers. The accidents have been classified by severity of impact into mild, moderate, severe, extreme, giving four 2×2 tables. Under the mild impacts, both p_{11} and p_{12} may be small and δ_1 also small, since injury rarely occurs with mild impact. Under extreme impact, p_{41} and p_{42} may both be close to 100%, making δ_4 also small. The large δ 's may occur in the two middle tables where the p 's are nearer 50%.

In applications of this type, two mathematical models have been used to describe how δ_i may be expected to change as p_{i2} changes. One model supposes that the difference between the two populations is constant on a *logit* scale. The logit of a proportion p is $\log_e(p/q)$. A constant difference on the logit scale means that $\log_e(p_{11}/q_{11}) - \log_e(p_{12}/q_{12})$ is constant as p_{12} varies. The second model postulates that the difference is constant on a *normal deviate* (Z) scale. The value of Z corresponding to any proportion p is such that the area of a standard normal curve to the

left of Z is p . For instance, $Z = 0$ for $p = 0.5$, $Z = 1.282$ for $p = 0.9$, $Z = -1.282$ for $p = 0.1$.

To illustrate the meaning of a constant difference on these transformed scales, table 9.14.2 shows the size of difference on the original percentage scale that corresponds to a constant difference on (a) the logit scale (b) the normal deviate scale. The size of the difference was chosen to equal 20% at $p_2 = 50\%$. Note that (i) the differences diminish towards both ends of the p scale as in the seat belts example, (ii) the two transformations do not differ greatly.

TABLE 9.14.2
SIZE OF DIFFERENCE $\delta = p_1 - p_2$ FOR A RANGE OF VALUES OF p_2

$p_2\%$	1	5	10	30	50	70	90	95	99
Constant logit	2.6	8.1	12.4	20.0	20.0	15.3	6.4	3.5	0.8
Constant Z	1.3	6.0	10.6	20.0	20.0	14.5	5.5	2.8	0.6

A test that gives appropriate weight to tables with large n_i and is sensitive if differences are constant on a logit or a Z scale was developed by Cochran (9). If \hat{p}_i is the combined percentage in the i th table, and

$$w_i = n_{i1}n_{i2}/(n_{i1} + n_{i2}) : d_i = \hat{p}_{i1} - \hat{p}_{i2},$$

we compute

$$\Sigma w_i d_i / \sqrt{\Sigma w_i \hat{p}_i \hat{q}_i}$$

and refer to the normal table. For the data in table 9.14.1 the computations are as follows (with the d_i in proportions to keep the numbers smaller).

Birth Order	w_i	d_i	$w_i d_i$	\hat{p}_i	$\hat{p}_i \hat{q}_i$	$w_i \hat{p}_i \hat{q}_i$
2	39.3	+0.040	+1.57	0.181	0.1482	5.824
3-4	27.3	+0.040	+1.09	0.372	0.2336	6.377
5+	21.1	+0.173	+3.65	0.477	0.2494	5.262
Sum			+6.31			17.463

The test criterion is $6.31 / \sqrt{17.463} = 1.51$. This agrees closely with the value 1.54 found by the $\Sigma \chi$ test, for which these tables are quite suitable.

There is another way of computing this test. In the i th table, let O_i be the observed number of Problems losses and E_i the expected number under H_0 . For birth order 2 (table 9.14.1), $O_i = 20$, $E_i = (30)(102)/166$

TABLE 9.14.3
THE MANTEL-HAENSZEL TEST FOR THE INFANT LOSS DATA IN TABLE 9.14.1

Birth Order	O_i	E_i	$n_{i1}n_{i2}c_{i1}c_{i2}/n_i^2(n_i - 1)$
2	20	18.43	5.858
3-4	26	24.90	6.426
5+	27	23.36	5.321
Sum	73	66.69	17.605

$$Z = (73 - 66.69 - \frac{1}{2})/\sqrt{17.605} = 1.38$$

= 18.43. Then $(O_1 - E_1) = +1.57$, which is the same as w_1d_1 . This result may be shown by algebra to hold in any 2×2 table. The criterion can therefore be written

$$\Sigma(O_i - E_i)/\sqrt{\Sigma w_i \hat{p}_i \hat{q}_i}$$

This form of the test has been presented by Mantel and Haenszel (28, 29), with two refinements that are worthwhile when the n 's are small. First, the variance of $w_i d_i$ or $(O_i - E_i)$ on H_0 is not $w_i \hat{p}_i \hat{q}_i$ but the slightly larger quantity $n_{i1}n_{i2}\hat{p}_i\hat{q}_i/(n_{i1} + n_{i2} - 1)$. If the margins of the 2×2 table are n_{i1} , n_{i2} , c_{i1} , and c_{i2} , this variance can be computed as

$$n_{i1}n_{i2}c_{i1}c_{i2}/n_i^2(n_i - 1), \quad (n_i = n_{i1} + n_{i2}),$$

a form that is convenient in small tables.

Secondly, a correction for continuity can be applied by subtracting $1/2$ from the absolute value of $\Sigma(O_i - E_i)$. This version of the test is shown in table 9.14.3. The correction for continuity makes a noticeable difference even with samples of this size.

The analysis of proportions is discussed further in sections 16.8-16.12.

REFERENCES

1. E. W. LINDSTROM. *Cornell Agric. Exp. Sta.*, Memoir 13 (1918).
2. A. W. F. EDWARDS. *Ann. Hum. Gen.*, 24:309 (1960).
3. J. H. EDWARDS. *Ann. Hum. Gen.*, 25:89 (1961).
4. C. W. LEGGATT. *Comptes rendus de l'association internationale d'essais de semences*, 5:27 (1935).
5. D. J. CAFFEY and C. E. SMITH. Bureau of Entomology and Plant Quarantine, USDA (Baton Rouge) (1934).
6. W. G. COCHRAN. *Ann. Math. Statist.*, 23:315 (1952).
7. I. M. CHAKRAVARTI and C. R. RAO. *Sankhyā*, 21:315 (1959).
8. W. G. COCHRAN. *J. R. Statist. Soc. Suppl.*, 3:49 (1936).
9. W. G. COCHRAN. *Biometrics*, 10:417 (1954).
10. G. W. SNEDECOR and M. R. IRWIN. *Iowa State Coll. J. Sci.*, 8:75 (1933).
11. R. C. LEWONTIN and J. FELSENSTEIN. *Biometrics*, 21:19 (1965).
12. J. B. S. HALDANE. *Biometrika*, 33:234 (1943-46).
13. J. O. IRWIN and E. A. CHEESEMAN. *J. R. Statist. Soc. Suppl.* 6:174 (1939).
14. L. C. BURNETT. M.S. Thesis. Iowa State College (1906).
15. G. C. DECKER and F. ANDRE. *Iowa State J. Sci.*, 10:403 (1936).

16. A. W. KIMBALL. *Biometrics*, 10:452 (1952).
17. D. J. BARTHOLOMEW. *Biometrika*, 46:328 (1959).
18. F. YATES. *Biometrika*, 35:176 (1948).
19. P. ARMITAGE. *Biometrics*, 11:375 (1955).
20. K. PEARSON. *Biometrika*, 5:105 (1905-06).
21. I. D. J. BROSS. *Biometrics*, 14:18 (1958).
22. J. IPSEN. *Biometrics*, 12:465 (1955).
23. R. A. FISHER. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh (1941).
24. R. P. ABELSON and J. W. TUKEY. *Proc. Soc. Statist. Sect. Amer. Statist. Ass.* (1959).
25. R. A. FISHER. *Ann. Sci.*, 1:117 (1936).
26. N. V. STRAND and R. J. JESSEN. *Iowa Agr. Exp. Stat. Res. Bul.* 315 (1943).
27. P. H. LESLIE. *Biometrics*, 7: 283 (1951).
28. N. MANTEL and W. HAENSZEL. *J. Nat. Cancer Inst.*, 22:719 (1959).
29. N. MANTEL. *J. Amer. Statist. Ass.*, 58:690 (1963).

One-way classifications.

Analysis of variance

10.1—Extension from two samples to many. Statistical methods for two independent samples were presented in chapter 4, but the needs of the investigator are seldom confined to the comparison of two samples only. For attribute data, the extension to more than two samples was made in the preceding chapter. We are now ready to do the same for measurement data.

First, recall the analysis used in the comparison of two samples. In the numerical example (section 4.9, p. 102), the comb weights of two samples of 11 chicks were compared, one sample having received sex hormone A, the other sex hormone C. Briefly, the principal steps in the analysis were as follows: (i) the mean comb weights \bar{X}_1, \bar{X}_2 were computed, (ii) the within-sample sum of squares of deviations Σx^2 , with 10 *df.*, was found for each sample, (iii) a pooled estimate s^2 of the within-sample variance was obtained by adding the two values of Σx^2 and dividing by the sum of the *df.*, 20, (iv) the standard error of the mean difference, $\bar{X}_1 - \bar{X}_2$, was calculated as $\sqrt{(2s^2/n)}$, where $n = 11$ is the size of each sample, (v) finally, a test of the null hypothesis $\mu_1 = \mu_2$ and confidence limits for $\mu_1 - \mu_2$ were given by the result that the quantity

$$\{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)\} / \sqrt{(2s^2/n)}$$

follows the *t*-distribution with 20 *df.*

In the next section we apply this method to an experiment with four treatments, i.e., four independent samples.

10.2—An experiment with four samples. During cooking, doughnuts absorb fat in various amounts. Lowe (1) wished to learn if the amount absorbed depends on the type of fat used. For each of four fats, six batches of doughnuts were prepared, a batch consisting of 24 doughnuts. The data in table 10.2.1 are the grams of fat absorbed per batch, coded by deducting 100 grams to give simpler figures. Data of this kind are called a *single* or *one-way* classification, each fat representing one class.

Before beginning the analysis, note that the totals for the four fats differ substantially, from 372 for fat 4 to 510 for fat 2. Indeed, there is a

TABLE 10.2.1
GRAMS OF FAT ABSORBED PER BATCH (MINUS 100 GRAMS)

Fat	1	2	3	4	Total
	64	78	75	55	
	72	91	93	66	
	68	97	78	49	
	77	82	71	64	
	56	85	63	70	
	95	77	76	68	
ΣX	432	510	456	372	1,770 = G
\bar{X}	72	85	76	62	295
ΣX^2	31,994	43,652	35,144	23,402	134,192
$(\Sigma X)^2/n$	31,104	43,350	34,656	23,064	132,174
Σx^2	890	302	488	338	2,018
$d.f.$	5	5	5	5	20

$$\text{Pooled } s^2 = 2,018/20 = 100.9$$

$$s_B = \sqrt{(2s^2/n)} = \sqrt{(2)(100.9)/6} = 5.80$$

clear separation between the individual results for fats 4 and 2, the highest value given by fat 4 being 70, while the lowest for fat 2 is 77. Every other pair of samples, however, shows some overlap.

Proceeding as in the case of two samples, we calculate for each sample the mean \bar{X} and the sum of squares of deviations Σx^2 , as shown under table 10.2.1. We then form a pooled estimate s^2 of the within-sample variance. Since each sample provides 5 $d.f.$ for Σx^2 , the pooled $s^2 = 100.9$ has 20 $d.f.$ This pooling involves, of course, the assumption that the variance between batches is the same for each fat. The standard error of the mean of any batch is $\sqrt{s^2/6} = 4.10$ grams.

Thus far, the only new problem is that there are four means to compare instead of two. The comparisons that are of interest are not necessarily confined to the differences $\bar{X}_i - \bar{X}_j$ between pairs of means: their exact nature will depend on the questions that the experiment is intended to answer. For instance, if fats 1 and 2 were animal fats and fats 3 and 4 vegetable fats, we might be particularly interested in the difference $(\bar{X}_1 + \bar{X}_2)/2 - (\bar{X}_3 + \bar{X}_4)/2$. A rule for making planned comparisons of this nature is outlined in section 10.7, with further discussion in sections 10.8, 10.9.

Before considering the comparison of means, we present an alternative method of doing the preliminary calculations in this section. This method, of great utility and flexibility, is known as the *analysis of variance* and was developed by Fisher in the 1920's. The analysis of variance performs two functions:

1. It is an elegant and slightly quicker way of computing the pooled s^2 .
2. In a single classification this advantage in speed is minor, but in the

more complex classifications studied later, the analysis of variance is the only simple and reliable method of determining the appropriate pooled error variance s^2 .

2. It provides a new test, the F -test. This is a single test of the null hypothesis that the population means $\mu_1, \mu_2, \mu_3, \mu_4$, for the four fats are identical. This test is often useful in a preliminary inspection of the results and has many subsequent applications.

EXAMPLE 10.2.1—Here are some data selected for easy computation. Calculate the pooled s^2 and state how many df it has.

Sample number			
1	2	3	4
11	13	21	10
4	9	18	4
6	14	15	19

Ans. $s^2 = 21.5$, with 8 df

10.3—The analysis of variance. In the doughnut example, suppose for a moment that there are no differences between the average amounts absorbed for the four fats. In this situation, all 24 observations are distributed about a common mean μ with variance σ^2 .

The analysis of variance develops from the fact that we can make three different estimates of σ^2 from the data in table 10.2.1. Since we are assuming that all 24 observations come from the same population, we can compute the total sum of squares of deviations for the 24 observations as

$$\begin{aligned} & 64^2 + 72^2 + 68^2 + \dots + 70^2 + 68^2 - (1770)^2/24 \\ & = 134,192 - 130,538 = 3654 \end{aligned} \quad (10.3.1)$$

This sum of squares has 23 df . The mean square, $3654/23 = 158.9$, is the first estimate of σ^2 .

The second estimate is the pooled s^2 already obtained. Within each fat, we computed the sum of squares between batches (890, 302, etc.), each with 5 df . These sums of squares were added to give

$$890 + 302 + 488 + 338 = 2018 \quad (10.3.2)$$

This quantity is called the sum of squares *between batches within fats*, or more concisely the sum of squares *within fats*. The sum of squares is divided by its df , 20, to give the second estimate, $s^2 = 2,018/20 = 100.9$.

For the third estimate, consider the means for the four fats, 72, 85, 76, and 62. These are also estimates of μ , but have variances $\sigma^2/6$, since they are means of samples of 6. Their sum of squares of deviations is

$$72^2 + 85^2 + 76^2 + 62^2 - (295)^2/4 = 272.75$$

with 3 *d.f.* The mean square, $272.75/3$, is an estimate of $\sigma^2/6$. Consequently, if we multiply by 6, we have the third estimate of σ^2 . We shall accomplish this by multiplying the sum of squares by 6, giving

$$6\{72^2 + 85^2 + 76^2 + 62^2 - (295)^2/4\} = 1636 \quad (10.3.3)$$

the mean square being $1636/3 = 545.3$.

Since the total for any fat is six times the fat means, this sum of squares can be computed from the fat totals as

$$\begin{aligned} & \frac{432^2 + 510^2 + 456^2 + 372^2}{6} - \frac{(1770)^2}{24} \\ &= 132,174 - 130,538 = 1636 \end{aligned} \quad (10.3.4)$$

To verify this alternative form of calculation, note that $432^2/6 = (6 \times 72)^2/6 = 6(72)^2$, while $(1770)^2/24 = (6 \times 295)^2/24 = 6(295)^2/4$. This sum of squares is called the sum of squares *between fats*.

Now list the *d.f.* and the sums of squares in (10.3.3), (10.3.2), and (10.3.1) as follows:

Source of Variation	Degrees of Freedom	Sum of Squares
Between fats	3	1,636
Between batches within fats	20	2,018
Total	23	3,654

Notice a new and important result: the *d.f.* and the sums of squares for the two components (between fats and within fats) add to the corresponding total figures. These results hold in any single classification. The result for the *d.f.* is not hard to verify. With a classes and n observations per class, the *d.f.* are $(a - 1)$ for Between fats, $a(n - 1)$ for Within fats, and $(an - 1)$ for the total. But

$$(a - 1) + a(n - 1) = a - 1 + an - a = an - 1$$

The result for the sums of squares follows from an algebraic identity (example 10.3.5). Because of this relation, the standard practice in the analysis of variance is to compute only the total sum of squares and the sum of squares Between fats. The sum of squares Within fats, leading to the pooled s^2 , is obtained by subtraction.

Table 10.3.1 shows the usual analysis of variance table for the doughnut data, with general computing instructions for a classes (fats) with n observations per class. The symbol T denotes a typical class total, while $G = \Sigma T = \Sigma \Sigma X$ (summed over both rows and columns) is the grand total. The first step is to calculate the *correction for the mean*,

$$C = G^2/an = (1770)^2/24 = 130,538$$

This is done because C occurs both in formula (10.3.1) for the total sum of squares and in formula (10.3.4) for the sum of squares between fats. The remaining steps should be clear from table 10.3.1.

TABLE 10.3.1
FORMULAS FOR CALCULATING THE ANALYSIS OF VARIANCE TABLE
(ILLUSTRATED BY THE DOUGHNUT DATA)

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Between classes (fats)	$a - 1 = 3$	$(\Sigma T^2/n) - C = 1,636$	545.3
Within classes (fats)	$a(n - 1) = 20$	Subtract = 2,018	100.9
Total	$an - 1 = 23$	$\Sigma \Sigma X^2 - C = 3,654$	

Since the analysis of variance table is unfamiliar at first, the beginner should work a number of examples. The role of the mean square between fats, which is needed for the F -test, is explained in the next section.

EXAMPLE 10.3.1—From the formulas in table 10.3.1, compute the analysis of variance for the simple data in example 10.2.1. Verify that you obtain 21.5 for the pooled s^2 , as found by the method of example 10.2.1.

Source of Variation	d.f.	Sum of Squares	Mean Square
Between samples	3	186	62.0
Within samples	8	172	21.5
Total	11	358	32.5

EXAMPLE 10.3.2—As part of a larger experiment (2), three levels of vitamin B_{12} were compared, each level being fed to three different pigs. The average daily gains in weight of the pigs (up to 75 lbs. live weight) were as follows:

Level of B_{12} (mg /lb ration)		
5	10	20
1.52	1.63	1.44
1.56	1.57	1.52
1.54	1.54	1.63

Analyze the variance as follows

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Between levels	2	0.0042	0.0021
Within levels	6	0.0232	0.0039
Total	8	0.0274	0.0034

Hint If you subtract 1.00 from each gain (or 1.44 if you prefer it) you will save time. Subtraction of a common figure from every observation does not alter any of the results in the analysis of variance table.

EXAMPLE 10.3.3—In table 9.4.1 there were recorded the number of loopers (insect larvae) on 50 cabbage plants per plot after the application of five treatments to each of four plots. The numbers were:

Treatment				
1	2	3	4	5
11	6	8	14	7
4	4	6	27	4
4	3	4	8	9
5	6	11	18	14

With counts like these, there is some question whether the assumptions required for the analysis of variance are valid. But for illustration, analyze the variance as follows:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Between treatments	4	359.30	89.82
Within treatments	15	311.25	20.75
Total	19	670.55	

EXAMPLE 10.3.4—The percentage of clean wool in seven bags was estimated by taking three batches at random from each bag. The percentages of clean wool in the batches were as follows:

Bag Number						
1	2	3	4	5	6	7
41.8	33.0	38.5	43.7	34.2	32.6	36.2
38.9	37.5	35.9	38.9	38.6	38.4	33.4
36.1	33.1	33.9	36.3	40.2	34.8	37.9

Calculate the mean squares for bags (11.11) and batches within bags (8.22).

EXAMPLE 10.3.5—To prove the result that the sums of squares within and between classes add to the total sum of squares, we use a notation that has become common for this type of data. Let X_{ij} be the observation for the j th member of the i th class. $X_{i\cdot}$ is the total of the i th class and $X_{\cdot\cdot}$ the grand total.

The sum of squares within the i th class is

$$\sum_{j=1}^n X_{ij}^2 - X_{i\cdot}^2/n$$

On adding this quantity over all classes to get the numerator of the pooled s^2 , we obtain, for the sum of squares within classes

$$\sum_{i=1}^a \sum_{j=1}^n X_{ij}^2 - \sum_{i=1}^a X_{i\cdot}^2/n \quad (1)$$

The sum of squares between classes is computed as

$$\sum_{i=1}^a X_{i\cdot}^2/n - X_{\cdot\cdot}^2/an \quad (2)$$

The sum of (1) and (2) gives

$$\sum_{i=1}^a \sum_{j=1}^n X_{ij}^2 - X^2/an$$

But this is the total sum of squares of deviations for the overall mean

10.4—Effect of differences between the population means. If the population means for the four fats are identical, we have seen that the mean square between fats, 545.3, and the mean square within fats, 100.9, are both estimates of the population variance σ^2 . What happens when the population means are different? In order to illustrate from a simple example in which you can easily verify the calculations, we drew (using a table of random normal deviates) six observations normally distributed with population mean $\mu = 5$ and $\sigma = 1$. These were arranged in three sets of two observations, to simulate an experiment with $a = 3$ treatments and $n = 2$ observations per treatment.

TABLE 10 4.1
A SIMULATED EXPERIMENT WITH THREE TREATMENTS AND
TWO OBSERVATIONS PER TREATMENT

Data			Analysis of Variance				
Case I.	$\mu_1 = \mu_2 = \mu_3 = 5$			<i>df</i>	<i>S.S.</i>	<i>M.S.</i>	
1	Treatment			Treatments	2	1.66	0.83
	2	3	Error	3	3.37	1.12	
	4.6	3.3	Total	5	5.03		
	5.2	4.7					
	9.8	8.0					
10.5							
Case II.	$\mu_1 = 4, \mu_2 = 5, \mu_3 = 7$			<i>df</i>	<i>S.S.</i>	<i>M.S.</i>	
1	Treatment			Treatments	2	14.53	7.26
	2	3	Error	3	3.37	1.12	
	3 6	3.3	Total	5	17.90		
	4 2	4 7					
	7.8	8 0					
14 5							
Case III.	$\mu_1 = 3, \mu_2 = 5, \mu_3 = 9$			<i>df</i>	<i>S.S</i>	<i>M.S.</i>	
1	Treatment			Treatments	2	46.06	23.03
	2	3	Error	3	3 37	1.12	
	2.6	3.3	Total	5	49.43		
	3.2	4.7					
	5 8	8.0					
18.5							

The data and the analysis of variance appear as Case I at the top of table 10.4.1. In the analysis of variance table, the Between classes sum of squares is labeled Treatments, and the Within classes sum of squares is labeled Error. This terminology is common in planned experiments. The mean squares, 0.83 for Treatments and 1.12 for Error, are both estimates of $\sigma^2 = 1$.

In Case II we subtracted 1 from each observation for treatment 1 and added 2 to each observation for treatment 3. This simulates an experiment with real differences in the effects of the treatments, the population means being $\mu_1 = 4$, $\mu_2 = 5$, $\mu_3 = 7$. In the analysis of variance, notice that the Error sum of squares and mean square are unchanged. This should not be surprising, because the Error *S.S.* is the pooled Σx^2 within treatments, and subtracting any constant from all the observations in a treatment has no effect on Σx^2 . The Treatments mean square has, however, increased from 0.83 in Case I to 7.26 in Case II.

Case III represents an experiment with larger differences between treatments. Each original observation for treatment 1 was reduced by 2, and each observation for treatment 3 was increased by 4. The means are now $\mu_1 = 3$, $\mu_2 = 5$, $\mu_3 = 9$. As before, the Error mean square is unchanged. The Treatments mean square has increased to 23.03. Note that the samples for the three treatments have now moved apart, so that there is no overlap.

When the means μ_i differ, it can be proved that the Treatments mean square is an unbiased estimate of

$$\sigma^2 + n \sum_{i=1}^a (\mu_i - \bar{\mu})^2 / (a - 1) \quad (10.4.1)$$

In Case II, with $\mu_1 = 4$, 5, 7, $\Sigma(\mu_i - \bar{\mu})^2$ is 4.67, while n and $(a - 1)$ are both 2 and $\sigma^2 = 1$, so that (10.4.1) becomes $1 + 4.67 = 5.67$. Thus the Treatments mean square, 7.26, is an unbiased estimate of 5.67. If we drew a large number of samples and calculated the Treatments mean square for Case II for each sample, their average should be close to 5.67.

In Case III, $\Sigma(\mu_i - \bar{\mu})^2$ is 18.67, so that the Treatments mean square, 23.03, is an estimate of the population value 19.67.

10.5—The variance ratio, F . These results suggest that the quantity,

$$F = \frac{\text{Treatments mean square}}{\text{Error mean square}} = \frac{\text{Mean square between classes}}{\text{Mean square within classes}},$$

should be a good criterion for testing the null hypothesis that the population means are the same in all classes. The value of F should be around 1 when the null hypothesis holds, and should become large when the μ_i differ substantially. The distribution was first tabulated by Fisher in the form $z = \log_e \sqrt{F}$. In honor of Fisher, the criterion was named F by Snedecor (3). Fisher and Yates (4) designate F as the *variance ratio*.

In Case I, F is $0.83/1.12 = 0.74$. In Case II, F increases to $7.26/1.12 = 6.48$ and in Case III to $23.03/1.12 = 20.56$. When you have learned

how to read the F -table, you will find that in Case II, F , which has 2 and 3 degrees of freedom, is significant at the 10% level but not at the 5% level. In Case III, F is significant at the 5% level.

To give some idea of the distribution of F when the null hypothesis holds, a sampling experiment was conducted. Sets of 100 observations were drawn at random from the table of pig gains (table 3.2.1, p. 67), which simulates a normal population with $\mu = 30$, $\sigma = 10$. Each set was divided into $a = 10$ classes, each with $n = 10$ observations. The F ratio therefore has 9 *d.f.* in the numerator and 90 *d.f.* in the denominator.

TABLE 10.5.1
DISTRIBUTION OF F IN 100 SAMPLES FROM TABLE 3.2.1
(Degrees of freedom 9 and 90)

Class Interval	Frequency	Class Interval	Frequency
0. -0.24	7	1.50-1.74	5
0.25-0.49	16	1.75-1.99	2
0.50-0.74	16	2.00-2.24	4
0.75-0.99	26	2.25-2.49	2
1.00-1.24	11	2.50-2.74	2
1.25-1.49	8	2.75-2.99	1

Table 10.5.1 displays the sampling distribution of 100 values of F . One notices first the skewness; a concentration of small values and a long tail of larger values. Next, observe that 65 of the F are less than 1. If you remember that both terms of the ratio are estimates of σ^2 , you may be surprised that 1 is not the median. The mean, calculated as with grouped data, is 0.96: the theoretical mean is slightly greater than 1. Finally, 5% of the values lie beyond 2.25 and 1% beyond 2.75, so that these points are estimates of the 5% and 1% levels of the theoretical distribution.

Table A 14, Part I, contains the *theoretical* 5% and 1% points of F for convenient combinations of degrees of freedom. Across the top of the table is found f_1 degrees of freedom corresponding to the number of treatments (classes): $f_1 = a - 1$. At the left is f_2 , the degrees of freedom for individuals, $a(n - 1)$. Since the F -table is extensively used, table A 14, Part II, gives the 25%, 10%, 2.5%, and 0.5% levels.

To find the 5% and 1% points for the sampling experiment, look in the column headed by $f_1 = 9$ and down to the rows $f_2 = 80$ and 100. The required points are 1.98 and 2.62, halfway between those in the table. To be compared with these are the points experimentally obtained in table 10.5.1, 2.25 and 2.75; not bad estimates from a sample of 100 experiments. In order to check the sampling distribution more exactly, we went back to the original calculations and found 8% of the sample F 's beyond the 5% point and 2% beyond the 1%. This gives some idea of the variation to be encountered in sampling.

For the doughnut experiment, the hypothesis set up—that the batches

are random samples from populations with the same μ —may be judged by means of table A 14. From the analysis of variance in table 10.3.1,

$$F = 545.3/100.9 = 5.40$$

For $f_1 = 3$ and $f_2 = 20$, the 1% point in the new table is 4.94. Thus from the distribution specified in the hypothesis there is less than one chance in 100 of drawing a sample having a larger value of F . Evidently the samples come from populations with different μ 's. The conclusion is that the fats have different capabilities for being absorbed by doughnuts.

EXAMPLE 10.5.1—Four tropical feedstuffs were each fed to a lot of 5 baby chicks (9) The gains in weight were:

Lot	1	55	49	42	21	52
	2	61	112	30	89	63
	3	42	97	81	95	92
	4	169	137	169	85	154

Analyze the variance and test the equality of the μ . Ans. Mean squares: (i) lots, 8,745; (ii) chicks within lots, 722 $F = 12.1$ Since the sample F is far beyond the tabular 1% point, there is little doubt that the feedstuff populations have different μ 's.

EXAMPLE 10.5.2—In the wool data of example 10.3.4, test the hypothesis that the bags are all from populations with a common mean. Ans. $F = 1.35$, $F_{0.05} = 2.85$. There is not strong evidence against the hypothesis—the bags may all have the same percentage of clean wool

EXAMPLE 10.5.3—In the vitamin B₁₂ experiment of example 10.3.2, the mean gains for the three levels differ less than is to be expected from the mean square within levels. Although there is no reason for computing it, the value of F is 0.54. There is, of course, no evidence of differences among the μ .

EXAMPLE 10.5.4—In example 10.3.3, test the hypothesis that the treatments have no effect on the number of loopers. Ans. $F = 4.33$. What do you conclude?

10.6—Analysis of variance with only two classes. When there are only two classes, the F -test is equivalent to the t -test which we used in chapter 4 to compare the two means. With two classes, the relation $F = t^2$ holds. We shall verify this by computing the analysis of variance for the numerical example in table 4.9.1, p. 103. The pooled $s^2 = 16,220/20 = 811$, has already been computed in table 4.9.1. To complete the analysis of variance, compute the Between samples sum of squares. Since the sample totals were 1067 and 616, with $n = 11$, the sum of squares is,

$$\frac{(1067)^2 + (616)^2}{11} - \frac{(1683)^2}{22} = 9245.5 \quad (10.6.1)$$

With only two samples, this sum of squares is obtained more quickly as

$$\frac{(\sum X_1 - \sum X_2)^2}{2n} = \frac{(1067 - 616)^2}{(2)(11)} = 9245.5 \quad (10.6.2)$$

TABLE 10.6.1
ANALYSIS OF VARIANCE OF CHICK EXPERIMENT, TABLE 4.9.1

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Between samples	1	9,245.5	9,245.5
Within samples	20	16,220.0	811.0

$$F = 9,245.5/811.0 = 11.40 \quad \sqrt{F} = 3.38 = t$$

Table 10.6.1 shows the analysis of variance and the value of F , 11.40.

Note that $\sqrt{F} = 3.38$, the value of t found in table 4.9.1. Further, in the F table with $f_1 = 1$, the significance levels are the squares of those in the t table for the same f_2 . While it is a matter of choice which one is used, the fact that we are nearly always interested in the size and direction of the difference $(\bar{X}_1 - \bar{X}_2)$ favors the t -test.

EXAMPLE 10.6.1—Hansberry and Richardson (5) gave the percentages of wormy apples on two groups of 12 trees each. Group A , sprayed with lead arsenate, had 19, 26, 22, 13, 26, 25, 38, 40, 36, 12, 16, and 8% of apples wormy. Those of group B , sprayed with calcium arsenate and buffer materials, had 36, 42, 20, 43, 47, 49, 59, 37, 28, 49, 31, and 39% wormy. Compute the mean square Within samples, 111.41, with 22 $d.f.$; and that Between samples, 1650.04, with 1 $d.f.$ Then,

$$F = 1650.04/111.41 = 14.8$$

Next, test the significance of the difference between the sample means as in section 4.9. The value of t is $3.85 = \sqrt{14.8}$.

EXAMPLE 10.6.2—For $f_1 = 1$, $f_2 = 20$, verify that the 5% and 1% significance levels of F are the squares of those of t with 20 $d.f.$

EXAMPLE 10.6.3—Prove that the methods used in equations (10.6.1) and (10.6.2) in the text for finding the Between samples sum of squares, 9245.5, are equivalent.

EXAMPLE 10.6.4—From equation (10.6.2) it follows that $F = t^2$. For $F = (\Sigma X_1 - \Sigma X_2)^2 / 2ns^2$, while $t = (\bar{X}_1 - \bar{X}_2) / \sqrt{(2s^2/n)}$. Since $\bar{X}_1 = \Sigma X_1/n$, $\bar{X}_2 = \Sigma X_2/n$, we have $t = (\Sigma X_1 - \Sigma X_2) / \sqrt{(2ns^2)} = \sqrt{F}$.

10.7—Comparisons among class means. The analysis of variance is only the first step in studying the results. The next step is to examine the class means and the sizes of differences among them.

Often, particularly in controlled experiments, the investigator plans the experiment in order to estimate a limited number of specific quantities. For instance, in part of an experiment on sugar beet, the three treatments (classes) were: (i) mineral fertilizers (PK) applied in April one week before sowing, (ii) PK applied in December before winter ploughing, (iii) no minerals. The mean yields of sugar in cwt. per acre were as follows:

PK in April, $\bar{X}_1 = 68.8$, PK in December, $\bar{X}_2 = 66.8$, No PK , $\bar{X}_3 = 62.4$

The objective is to estimate two quantities:

Average effect of PK : $\frac{1}{2}(\bar{X}_1 + \bar{X}_2) - \bar{X}_3 = 67.8 - 62.4 = 5.4$ cwt.

April minus December application: $\bar{X}_1 - \bar{X}_2 = 2.0$ cwt.

A rule for finding standard errors and confidence limits of estimates of this type will now be given. Both estimates are linear combinations of the means, each mean being multiplied by a number. In the first estimate, the numbers are $1/2, 1/2, -1$. In the second, they are $1, -1, 0$, where we put 0 because \bar{X}_3 does not appear. Further, in each estimate, the sum of the numbers is zero. Thus,

$$\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right) + (-1) = 0 \quad : \quad (1) + (-1) + (0) = 0$$

Definition. Any linear combination,

$$L = \lambda_1 \bar{X}_1 + \lambda_2 \bar{X}_2 + \dots + \lambda_k \bar{X}_k,$$

where the λ 's are fixed numbers, is called a *comparison* of the treatment means if $\sum \lambda_i = 0$. The comparison may include all a treatment means, ($k = a$), or only some of the means ($k < a$).

Rule 10.7.1. The standard error of L is $\sqrt{\sum \lambda^2 (\sigma/\sqrt{n})}$, and the estimated standard error is $\sqrt{\sum \lambda^2 (s/\sqrt{n})}$, with degrees of freedom equal to those in s , where n is the number of observations in each mean \bar{X}_i .

In the example the value of s/\sqrt{n} was 1.37 with 24 *d.f.* Hence, for the average effect of *PK*, with $\lambda_1 = 1/2, \lambda_2 = 1/2, \lambda_3 = -1$, the estimated standard error is

$$\sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + (-1)^2} (1.37) = \sqrt{1.5} (1.37) = 1.68,$$

with 24 *d.f.* The value of t for testing the average effect of *PK* is $t = 5.4/1.68 = 3.2$, significant at the 1% level. Confidence limits (95%) are $5.4 \pm (2.06)(1.68)$, or 1.9 and 8.9 cwt. per acre.

For the difference between the April and December applications, with $\lambda_1 = 1, \lambda_2 = -1$, the estimated standard error is $\sqrt{2} (1.37) = 1.94$. The difference is not significant at the 5% level, the confidence limits being $2.0 \pm (2.06)(1.94)$, or -2.0 and $+6.0$.

In view of the importance of Rule 10.7.1, we shall sketch the proof of this result. Since the λ_i are fixed numbers, the population mean of L is

$$\mu_L = \lambda_1 \mu_1 + \lambda_2 \mu_2 + \dots + \lambda_k \mu_k$$

where μ_i is the population mean of \bar{X}_i . Hence,

$$L - \mu_L = \lambda_1 (\bar{X}_1 - \mu_1) + \lambda_2 (\bar{X}_2 - \mu_2) + \dots + \lambda_k (\bar{X}_k - \mu_k)$$

By definition, the variance of L is the average value of $(L - \mu_L)^2$ taken over the population. Now

$$(L - \mu_L)^2 = \sum_{i=1}^k \lambda_i^2 (\bar{X}_i - \mu_i)^2 + 2 \sum_{i=1}^k \sum_{j>i}^k \lambda_i \lambda_j (\bar{X}_i - \mu_i)(\bar{X}_j - \mu_j)$$

The average value of $(\bar{X}_i - \mu_i)^2$ over the population is of course the variance of \bar{X}_i . The average value of $(\bar{X}_i - \mu_i)(\bar{X}_j - \mu_j)$ is the quantity which we called the covariance of \bar{X}_i and \bar{X}_j (section 7.4, p. 181). This gives the general formula,

$$V(L) = \sum_{i=1}^k \lambda_i^2 V(\bar{X}_i) + 2 \sum_{i=1}^k \sum_{j>i}^k \lambda_i \lambda_j \text{cov}(\bar{X}_i, \bar{X}_j) \quad (10.7.1)$$

When the \bar{X}_i are the means of independent samples of size n , $V(\bar{X}_i) = \sigma^2/n$, and $\text{Cov.}(\bar{X}_i, \bar{X}_j) = 0$, giving

$$V(L) = (\sum \lambda_i^2) \sigma^2/n$$

in agreement with Rule 10.7.1.

When reporting the results of a series of comparisons, it is important to give the sizes of the differences, with accompanying standard errors or confidence limits. For any comparison of broad interest, it is likely that several experiments will be done, often by workers in different places. The best information on this comparison is a combined summary of the results of these experiments. In order to make this, an investigator needs to know the sizes of the individual results and their standard errors. If he is told merely that "the difference was not significant" or "the difference was significant at the 1% level," he cannot begin to summarize effectively.

For the example, a report might read as follows. "Application of mineral fertilizers produced a significant average increase in sugar of 5.4 cwt. per acre (± 1.68). The yield of the April application exceeded that of the December application by 2.0 cwt. (± 1.94), but this difference was not significant."

Comments: (i) Unless this is already clear, the report should state the the *amounts* of P and K that were applied; (ii) there is much to be said for presenting, in addition, a table of the treatment (class) means, with their standard error, ± 1.37 . This allows the reader to judge whether the general level of yield was unusual in any way, and to make other comparisons that interest him.

Further examples of planned comparisons appear in the next two chapters. Common cases are the comparison of a "no minerals" treatment with minerals applied in four different ways (section 11.3), the comparison of different levels of the same ingredient, usually at equal intervals, where the purpose is to fit a curve that describes the relation between yield and the amount of the ingredient (section 11.8), and factorial experimentation, which forms the subject of chapter 12.

Incidentally, when several different comparisons are being made, one or two of the comparisons may show significant effects even if the initial F -test shows non-significance.

The rule that a comparison L is declared significant at the 5% level if L/s_L exceeds $t_{0.05}$ is recommended for any comparisons that the experiment was designed to make. Sometimes, in examining the treatment means, we notice a combination which we did not intend to test but which seems unexpectedly large. If we construct the corresponding L , use of the t -test for testing L/s_L is invalid, since we selected L for testing solely because it looked large.

Scheffé (11) has given a general method that provides a conservative test in this situation. Declare L/s_L significant only if it exceeds $\sqrt{(a-1)F_{0.05}}$, where $F_{0.05}$ is the 5% level of F for degrees of freedom $f_1 = (a-1)$, $f_2 = a(n-1)$. In more complex experiments, f_2 is the number of error $d.f.$ provided by the experiment. Scheffé's test agrees with the t -test when $a = 2$, and requires a substantially higher value of L/s_L for significance when $a > 2$. It allows us to test any number of comparisons, picked out by inspection, with the protection that the probability of finding any erroneous significant result is at most 0.05.

EXAMPLE 10.7.1—In an experiment in which mangolds were grown on acid soil (6), part of the treatments were: (i) chalk, (ii) lime, both applied at the rate of 21 cwt. calcium oxide (CaO) per acre, and (iii) no liming. For good reasons, there were twice as many "no lime" plots as plots with chalk or with lime. Consequently, the comparisons of interest may be expressed algebraically as

$$\text{Effect of CaO: } \frac{1}{2}(\bar{X}_1 + \bar{X}_2) - \frac{1}{2}(\bar{X}_3 + \bar{X}_4)$$

where \bar{X}_3, \bar{X}_4 represent the two "no lime" classes.

Chalk minus lime: $\bar{X}_1 - \bar{X}_2$.

The mean yields were (tons per acre): chalk, 14.82; lime, 13.42; no lime, 9.74. The *s.e.* of any \bar{X}_i was ± 2.06 tons, with 25 *d.f.* Calculate the two comparisons and their standard errors, and write a report on the results. Ans. Effect of CaO, 4.38 ± 2.06 tons. Chalk minus lime, 1.40 ± 1.98 tons.

EXAMPLE 10.7.2—An experiment on sugar beet (7) compared times and methods of applying mixed artificial fertilizers (NPK). The mean yields of sugar (cwt. per acre) were as follows:

No Artificials	Artificials applied in:		
	Jan. (Ploughed)	Jan. (Broadcast)	Apr. (Broadcast)
38.7 \bar{X}_1	48.7 \bar{X}_2	48.8 \bar{X}_3	45.0 \bar{X}_4

Their *s.e.* was ± 1.22 , with 14 *d.f.* Calculate 95% confidence limits for the following comparisons:

Average effect of artificials $\frac{1}{3}(\bar{X}_2 + \bar{X}_3 + \bar{X}_4) - \bar{X}_1$

January minus April application: $\frac{1}{2}(\bar{X}_2 + \bar{X}_3) - \bar{X}_4$

Broadcast minus Ploughed in Jan.: $\bar{X}_3 - \bar{X}_2$

Ans.: (i) (5.8, 11.8); (ii) (0.6, 7.0); (iii) (-3.6, +3.8) cwt. per acre.

EXAMPLE 10.7.3—One can encounter linear combinations of the means that are not comparisons as we have defined them, but this seems to be rare. For instance, in early experiments on vitamin B_{12} , rats were fed on a B_{12} -deficient diet until they ceased to gain in weight. If we then compared a single and a double supplement of B_{12} , measuring the subsequent gains in weight produced, it might be reasonable to calculate $(\bar{X}_2 - 2\bar{X}_1)$, which should be zero if the gain in weight is proportional to the amount of B_{12} . Here $\lambda_1 + \lambda_2 \neq 0$. The formula for the standard error still holds. The *s.e.* is $\sqrt{3}\sigma/\sqrt{n}$ in this example.

10.8—Inspection of all differences between pairs of means. Often, the investigator has no specific comparisons, chosen in advance, that he proposes to make. Instead, he looks at all the means to see which

differences among them appear to be real. The most frequent example is when the treatments are qualitatively similar, as in tests on working gloves made by different manufacturers.

Taking the doughnut data from table 10.2.1 as an illustration, the means for the four fats (arranged in increasing order) are as follows:

TABLE 10.8.1

Fat	4	1	3	2	LSD	D
Mean grams absorbed	62	72	76	85	12.1	16.2

The standard error of the difference between two means, $\sqrt{(2s^2/n)}$, is ± 5.80 , with 20 *d.f.* (table 10.2.1). The 5% value of *t* with 20 *d.f.* is 2.086. Hence, the difference between a specific pair of means is significant at the 5% level if it exceeds $(2.086)(5.8) = 12.1$.

The highest mean, 85 for fat 2, is significantly greater than the means 72 for fat 1 and 62 for fat 4. The mean 76 for fat 3 is significantly greater than the mean 62 for fat 4. None of the other three differences between pairs reaches 12.1. The quantity 12.1 which serves as a criterion is called the *Least Significant Difference (LSD)*. Similarly, 95% confidence limits for the population difference between any pair of means are given by adding ± 12.1 to the observed difference.

Objections to indiscriminate use of the *LSD* in significance tests have been raised for many years. Suppose that all the population means μ_i are equal, so that there are no real differences. With five types of gloves, for instance, there are ten possible comparisons between pairs of means. The probability that at least one of the ten exceeds the *LSD* is bound to be greater than 0.05: it can be shown to be about 0.29. With ten means (45 comparisons among pairs) the probability of finding at least one significant difference is about 0.63 and with 15 means it is around 0.83.

When the μ_i are all equal, the *LSD* method still has the basic property of a test of significance, namely that about 5% of the tested differences will erroneously be declared significant. The trouble is that when many differences are tested, some that appear significant are almost certain to be found. If these are the ones that are reported and attract attention, the test procedure loses its valuable property of protecting the investigator against making erroneous claims.

Commenting on this issue, Fisher (8) wrote: "When the *z* test (i.e., the *F*-test) does not demonstrate significance, much caution should be used before claiming significance for special comparisons." In line with this remark, investigators are sometimes advised to use the *LSD* method only if *F* is significant.

Among other proposed methods, perhaps the best known is one which replaces the *LSD* by a criterion based on the tables of the Studentized Range, $Q = (\bar{X}_{\max} - \bar{X}_{\min})/s_{\bar{Y}}$. Table A 15 gives the upper 5% levels

of Q , i.e., the value exceeded in 5% of experiments. This value depends on the number of means, a , and the number f of $d.f.$ in s_X . Having read $Q_{0.05}$ from table A 15, we compute the difference D between two means that is required for 5% significance as $Q_{0.05} s_X$.

For the doughnuts, $a = 4$, $f = 20$, we find $Q_{0.05} = 3.96$. Hence $D = Q_{0.05} s_X = (3.96)(4.1) = 16.2$. Looking back at table 10.8.1, only the difference between fats 2 and 4 is significant with this criterion. When there are only two means, the Q method becomes identical with the LSD method. Otherwise Q requires a larger difference for significance than the LSD .

The Q method has the property that if we test some or all of the differences between pairs of means, the probability that no erroneous claim of significance will be made is ≥ 0.95 . Similarly, the probability that all the confidence intervals $(\bar{X}_i - \bar{X}_j) \pm D$ will correctly include the difference $\mu_i - \mu_j$ is 0.95. The price paid for this increased protection is, of course, that fewer differences $\mu_i - \mu_j$ that are real will be detected and that confidence intervals are wider.

EXAMPLE 10.8.1—In Case III of the constructed example in table 10.4.1, with $\mu_1 = 3$, $\mu_2 = 5$, $\mu_3 = 9$, the observed means are $\bar{X}_1 = 2.9$, $\bar{X}_2 = 4.0$, $\bar{X}_3 = 9.25$, with $s.e. = \sqrt{(s^2/n)} = 0.75$ (3 $d.f.$). Test the three differences by (i) the LSD test, (ii) the Q test. Construct a confidence interval for each difference by each method. (iii) Do all the confidence intervals include $(\mu_i - \mu_j)$? Ans. (i) $LSD = 3.37$. \bar{X}_3 significantly greater than \bar{X}_2 and \bar{X}_1 . (ii) Required difference = 4.43. Same significant differences. (iii) Yes.

EXAMPLE 10.8.2—In example 10.5.1, the mean gains in weight of baby chicks under four feeding treatments were $\bar{X}_1 = 43.8$, $\bar{X}_2 = 71.0$, $\bar{X}_3 = 81.4$, $\bar{X}_4 = 142.8$ while $\sqrt{(s^2/n)} = 12.0$ with 16 $d.f.$ Compare the means by the LSD and the Q methods. Ans. Both methods show that \bar{X}_4 differs significantly from any other mean. The LSD method gives \bar{X}_3 significantly greater than \bar{X}_1 .

Hartley (30) showed that a sequential variant of the Q method, originally due to Newman (10) and Keuls (31), gives the same type of protection and is more powerful; that is, the variant will detect real differences more frequently than the original Q method.

Arrange the means in ascending order. For the doughnut fats, these means are as follows:

Fat	4	1	3	2	$S.D$
	62	72	76	85	± 4.10 (20 $d.f.$)

As before, first test the extreme difference, fat 2 - fat 4 = 23, against $D = 16.2$. Since the difference exceeds D , proceed to test fat 2 - fat 1 = 13 and fat 3 - fat 4 = 14 against the D value for $a = 3$, because these comparisons are differences between the highest and lowest of a group of three means. For $a = 3$, $f = 20$, Q is 3.58, giving $D = (3.58)(4.10) = 14.7$. Both the differences, 13 and 14, fall short of D . Consequently we stop, the difference between fats 2 and 4 is the only significant difference in the

experiment. If fat 3 – fat 4 had been, say, 17, we would have declared this difference significant and next tested fat 3 – fat 1 and fat 1 – fat 4 against the D value for $\alpha = 2$.

Whenever the highest and lowest of a group of means are found not significantly different in this method, we declare that none of the members of this group is distinguishable. This rule avoids logical contradictions in the conclusions. The method is called *sequential* because the testing follows a prescribed order or sequence.

Since protection against false claims of significance is obtained by decreasing the ability to detect real differences, a realistic choice among these methods requires a judgment about the relative seriousness of the two kinds of mistake. Duncan (32) has examined the type of policy that emerges if the investigator assigns relative costs to (i) declaring a significant result when the true difference is zero, (ii) declaring non-significance when there is a true difference, (iii) declaring a significant result in the wrong direction. His policy is designed to minimize the average cost of mistakes in such verdicts of significance or non-significance. These costs are not necessarily monetary but might be in terms of utility or equity. His optimum policy resembles an *LSD* rule with two notable differences. In its simplest form, which applies when the number of treatments exceeds 15 and $d.f.$ in s exceed 30, a difference between two means is declared significant if it exceeds $s_{D'} t_{\alpha} \sqrt{F/(F-1)}$. The quantity t_{α} (not Student's t) depends on the relative costs assigned to wrong verdicts of significance or non-significance. If F is large, indicating that there are substantial differences among the population means of the treatments, $\sqrt{F/(F-1)}$ is nearly 1. The rule then resembles a simple *LSD* rule, but with the size of the *LSD* determined by the relative costs. As F approaches 1, suggesting that differences among treatment means are in general small, the difference required for significance becomes steadily larger, leading to greater caution in declaring differences significant. The F -value given by the experiment enters into the rule because F provides information as to whether real differences among treatment means are likely to be large or small. In Duncan's method, the investigator may also build into the rule his *a priori* judgment on this point.

In a large sampling experiment with four treatments, Balaam (33) compared (i) the *LSD* method, (ii) the revised *LSD*-method in which no significant differences are declared unless F is significant, (iii) the Newman-Keuls method (as well as other methods). Various sets of values were assigned to the population means μ_i , including a set in which all μ_i were equal. For each pair of means, a test procedure received a score of +1 if it ranked them correctly, a score 0 if it declared a significant difference when $\mu_i = \mu_j$ or found no difference when $\mu_i \neq \mu_j$, and a score -1 if it ranked the means in the wrong order. These scores were added over the six pairs of means.

When all μ_i were equal, the average scores were: *LSD*, 5.76; Revised *LSD*, 5.91; *NK*, 5.94. With three means equal, so that three of the six differences between pairs were equal and three unequal, average scores

were: *LSD*, 3.80; Revised *LSD*, 3.57; *NK*, 3.51. With more than three inequalities between pairs, average scores were: *LSD*, 1.92; Revised *LSD*, 1.73; *NK*, 1.63. To sum up for this section, no method is uniformly best. In critical situations, try to judge the relative costs of the two kinds of mistakes and be guided by these costs. For routine purposes, thoughtful use of either the *LSD* or the Newman-Keuls method should be satisfactory. Remember also Scheffé's test (p. 271) for a comparison that is picked out just because it looks large.

10.9—Shortcut computation using ranges. An easy method of testing all comparisons among means is based on the ranges of the samples (13). In the doughnut experiment, table 10.2.1, the four ranges are 39, 20, 30, 21; the sum is 110. This sum of ranges is multiplied by a factor taken from table 10.9.1. In the column for $a = 4$ and the row for $n = 6$, take the factor 0.95. Then

$$D' = \frac{(\text{Factor})(\text{Sum of Ranges})}{n} = \frac{(0.95)(110)}{6} = 17.4$$

D' is used like the D in the Q -test of the foregoing section. Comparing it with the six differences among treatments, we conclude, as before, that only the largest difference, 23, is significant.

TABLE 10.9.1
CRITICAL FACTORS FOR ALLOWANCES, 5% RISK*

Sample Size, n	Number of Samples, a								
	2	3	4	5	6	7	8	9	10
2	3.43	2.35	1.74	1.39	1.15	0.99	0.87	0.77	0.70
3	1.90	1.44	1.14	.94	.80	.70	.62	.56	.51
4	1.62	1.25	1.01	.84	.72	.63	.57	.51	.47
5	1.53	1.19	.96	.81	.70	.61	.55	.50	.45
6	1.50	1.17	.95	.80	.69	.61	.55	.49	.45
7	1.49	1.17	.95	.80	.69	.61	.55	.50	.45
8	1.49	1.18	.96	.81	.70	.62	.55	.50	.46
9	1.50	1.19	.97	.82	.71	.62	.56	.51	.47
10	1.52	1.20	.98	.83	.72	.63	.57	.52	.47

* Extracted from a more extensive table by Kurtz, Link, Tukey, and Wallace (13).

EXAMPLE 10.9.1—Using the shortcut method, examine all differences in the chick experiment of example 10.5.1 (p. 267). Ans. $D' = 49$. Same conclusions as for the Q method in example 10.8.2.

10.10—Model I. Fixed treatment effects. It is time to make a more formal statement about the assumptions underlying the analysis of variance for single classifications. A notation common in statistical papers is to use the subscript i to denote the class, where i takes on the values 1, 2, . . . a . The subscript j designates the members of a class, j going from 1 to n .

Within class i , the observations X_{ij} are assumed normally distributed about a mean μ_i with variance σ^2 . The mean μ_i may vary from class to class, but σ^2 is assumed the same in all classes. We denote the mean of the a values of μ_i by μ , and write $\mu_i = \mu + \alpha_i$. It follows, of course, that $\sum \alpha_i = 0$. Mathematically, the model may be written:

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}; \quad i = 1 \dots a, \quad j = 1 \dots n, \quad \varepsilon_{ij} = \mathcal{N}(0, \sigma).$$

In words:

Any observed value is the sum of three parts: (i) an overall mean, (ii) a treatment or class deviation, and (iii) a random element from a normally distributed population with mean zero and standard deviation σ .

The artificial data in table 10.4.1 were made up according to this model. In Case II, with $\mu_i = 4, 5, 7$, we have $\mu = 16/3$, $\alpha_1 = -4/3$, $\alpha_2 = -1/3$, $\alpha_3 = +5/3$. The ε_{ij} were drawn from a table of normal deviates with $\sigma = 1$.

This model is often called model I, the *fixed effects model*. Its distinctive feature is that the effects of the treatments or classes, measured by the parameters α_i , are regarded as fixed but unknown quantities to be estimated.

10.11—Effects of errors in the assumptions. For the user of the analysis of variance, two relevant questions are: (i) Are the assumptions satisfied in my data? (ii) Does it make any difference if they are not satisfied?

Real data are seldom, if ever, exactly normally distributed. Often they exhibit some skewness; if symmetrical, they may have longer tails than the normal distribution. Three situations in which one should be on the lookout for non-normality are: (i) with small whole numbers, whose distribution may approximate the Poisson rather than the normal, (ii) with proportions or percentages that cover a range extending nearly to zero or 100%, and (iii) cases in which the treatments (or classes) produce multiplicative effects. Model I assumes that the effect of the i th class is to add α_i to any existing value. If, instead, the effect is to multiply the existing value by, say, 60%, the observations are likely to approximate a distribution called the *lognormal*. This is a skew distribution of values X such that $\log X$ is normally distributed.

In a single classification with equal n , various mathematical studies agree in showing that the F -test is little affected by moderate non-normality. However, with non-normal data, the variance σ_i^2 within a class is often related to the mean μ_i of the class. For the Poisson distribution, you may recall that $\sigma_i^2 = \mu_i$. With a proportion, the variance may behave like $\mu_i(1 - \mu_i)$, and with the lognormal distribution, σ_i^2 tends to vary as μ_i^2 . It follows that with non-normal data, the use of a pooled estimate of error s^2 in comparing pairs or subgroups of means can be seriously misleading. With two treatments A and B that produce small means, σ^2 might be about 20, while with C and D , which give large means, σ^2 is about 60. The pooled s^2 will be about 40. For comparing A with B , the pooled s^2 gives a t -value that is too small by a factor $\sqrt{2} = 1.41$,

while for comparing C with D , t is too large by a factor $\sqrt{3/2}$. Heterogeneous variance also occurs occasionally because some treatments by their nature produce erratic effects—sometimes they work well, sometimes not. Here there may be no clear relation between σ_i^2 and μ_i .

When comparing two classes, a safe rule is to calculate s^2 from the data for these two classes only. The disadvantage is that the number of *d.f.* is reduced (see also section 4.14). With a single erratic treatment (the i th), a pooled s^2 can be calculated and used for comparisons among the remaining treatments, and a separate s_i^2 for the erratic one. The *s.e.* of $(\bar{X}_i - \bar{X}_j)$ is estimated as

$$\sqrt{(s_i^2 + s^2)/n}$$

When the relation between σ_i^2 and μ_i is caused by non-normality, a knowledge of the type of data, plus a look at the relation between \bar{X}_i and R_i (the range within the class) helps in deciding whether the data are of the Poisson type ($R_i \propto \sqrt{\bar{X}_i}$), the quasi-binomial type ($R_i \propto \sqrt{\bar{X}_i(1 - \bar{X}_i)}$), or the lognormal type, $R_i \propto \bar{X}_i$. For these three types, transformations will be given later (sections 11.14–11.17) that bring the data closer to normality and often permit the use of a pooled error variance for all comparisons.

10.12—Samples of unequal sizes. In planned experiments, the samples from the classes are usually made of equal sizes, but in non-experimental studies the investigator may have little control over the sizes of the samples. As before, X_{ij} denotes the j th observation from the i th class. The symbol $X_{i.}$ denotes the class total of the X_{ij} , while $X_{..} = \Sigma X_{i.}$ is the grand total. The size of the sample in the i th class is n_i , and $N = \Sigma n_i$ is the total size of all samples. The correction for the mean is

$$C = X_{..}^2/N$$

Algebraic instructions for the *d.f.* and sums of squares in the analysis of variance appear in table 10.12.1.

TABLE 10.12.1
ANALYSIS OF VARIANCE WITH SAMPLES OF UNEQUAL SIZES

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Between classes	$a - 1$	$\sum \frac{X_{i.}^2}{n_i} - C$	s_c^2
Within classes	$N - a$	Subtract = $\Sigma \Sigma X_{ij}^2 - \sum \frac{X_{i.}^2}{n_i}$	s^2
Total	$N - 1$	$\Sigma \Sigma X_{ij}^2 - C$	

The F ratio, s_c^2/s^2 , has $(a - 1)$ and $(N - a)$ *d.f.* The *s.e.* of the difference between the i th and the k th class means, with $(N - a)$ *d.f.*, is

$$\sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_k} \right)}$$

The s.e. of the comparison $\Sigma \lambda_i \bar{X}_i$ is

$$\sqrt{s^2 \Sigma \frac{\lambda_i^2}{n_i}}$$

With unequal n_i , the F - and t -tests are more affected by non-normality and heterogeneity of variances than with equal n_i (14). Bear this in mind when starting to analyze the data.

EXAMPLE 10.12.1—The numbers of days survived by mice inoculated with three strains of typhoid organisms are summarized in the following frequency distributions. Thus, with strain 9D, 6 mice survived for 2 days, etc. We have $n_1 = 31$, $n_2 = 60$, $n_3 = 133$, $N = 224$. The purpose of the analysis is to estimate and compare the mean numbers of days to death for the three strains.

Since the variance for strain 9D looks much smaller than that for the other strains, it seems wise to calculate s_i^2 separately for each strain, rather than use a pooled s^2 from the analysis of variance. The calculations are given under the table.

Days to Death	Numbers of Mice Inoculated With Indicated Strain			Total
	9D	11C	DSC1	
2	6	1	3	10
3	4	3	5	12
4	9	3	5	17
5	8	6	8	22
6	3	6	19	28
7	1	14	23	38
8		11	22	33
9		4	14	18
10		6	14	20
11		2	7	9
12		3	8	11
13		1	4	5
14			1	1
Total	31	60	133	224
ΣX	125	442	1,037	1,604
ΣX^2	561	3,602	8,961	13,124
n_i	31	60	133	224
X_i	125	442	1,037	1,604
\bar{X}_i	4.03	7.37	7.80	
ΣX_i^2	561	3,602	8,961	13,124
X_i^2/n_i	504	3,256	8,085	
$\Sigma (X_{ij} - \bar{X}_i)^2$	57	346	876	
s_i^2	1.90	5.86	6.64	

The difference in mean days to death for strains 11C and 9D is 3.34 days, with

$$s.e. = \sqrt{\left\{ \frac{1.90}{31} + \frac{5.86}{60} \right\}} = \sqrt{0.1591} = \pm 0.399.$$

For strains DSC1 and 11C, the difference is 0.43 days ± 0.384

EXAMPLE 10.12.2—As an exercise, calculate the analysis of variance for the preceding data. Show that $F = 179.5/5.79 = 31.0$, $f = 2$ and 221. Show that if the pooled s^2 were used, the *s.e.* of the mean difference between strains 11C and 9D would be estimated as ± 0.532 instead of ± 0.399

10.13—Model II. Random effects. With some types of single classification data, the model used and the objectives of the analysis differ from those under model I. Suppose that we wish to determine the average content of some chemical in a large population or batch of leaves. We select a random sample of a leaves from the population. For each selected leaf, n independent determinations of the chemical content are made giving $N = an$ observations in all. The leaves are the classes, and the individual determinations are the members of a class.

In model II, the chemical content found for the j th determination from the i th leaf is written as

$$X_{ij} = \mu + A_i + \varepsilon_{ij}, \quad i = 1 \dots a, j = 1 \dots n, \quad (10.13.1)$$

where

$$A_i = \mathcal{N}(0, \sigma_A); \quad \varepsilon_{ij} = \mathcal{N}(0, \sigma)$$

The symbol μ is the mean chemical content of the population of leaves. This is the quantity to be estimated. The symbol A_i represents the difference between the chemical content of the i th leaf and the average content over the population. By including this term, we take account of the fact that the content varies from leaf to leaf. Every leaf in the population has its value of A_i , so that we may think of A_i as a random variable with a distribution over the population. This distribution has mean 0, since the A_i are defined as deviations from the population mean. In the simplest version of model II, it is assumed in addition that the A_i are normally distributed with standard deviation σ_A . Hence, we have written $A_i = \mathcal{N}(0, \sigma_A)$.

What about the term ε_{ij} ? This term is needed because:

- (i) the determination is subject to an error of measurement, and
- (ii) if the determination is made on a small piece of the leaf, its content may differ from that of the leaf as a whole. The ε_{ij} and the A_i are assumed independent. The further assumption $\varepsilon_{ij} = \mathcal{N}(0, \sigma)$ is often made.

There are some similarities and some differences between model II and model I. In model I

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \alpha_i \text{ fixed}, \quad \varepsilon_{ij} = \mathcal{N}(0, \sigma)$$

Note the following points:

(i) The α_i are fixed quantities to be estimated; the A_i are random variables. As will be seen, their variance σ_A^2 is often of interest.

(ii) The null hypothesis $\alpha_i = 0$ is identical with the null hypothesis $\sigma_A = 0$, since in this event all the A_i must be zero. Thus, the F test holds also in model II, being now a test of the null hypothesis $\sigma_A = 0$.

(iii) We saw (in section 10.4), that when the null hypothesis is false, the mean square between classes under model I is an unbiased estimate of

$$E(M.S. \text{ Between}) = \sigma^2 + n\sum\alpha_i^2/(a-1) \quad (10.13.2)$$

There is an analogous result for model II, the mean square estimating

$$E(M.S. \text{ Between}) = \sigma^2 + n\sigma_A^2 \quad (10.13.3)$$

Neither result requires the assumption of normality.

(iv) In drawing repeated samples under model I, we always draw from the same set of classes with the same α_i . Under model II, we draw a *new* random sample of a leaves. A consequence is that the general distributions of F (when the H_0 is false) differ. With model I, this distribution, the power function, is complicated: tables by Tang (15) and charts by Pearson and Hartley (16) are available. With model II, the probability that the observed variance ratio exceeds any value F_0 is simply the probability that the ordinary F exceeds $F_0/(1 + n\sigma_A^2/\sigma^2)$.

To turn to an example of model II, the data for calcium in table 10.13.1 come from a large experiment (17) on the precision of estimation of the chemical content of turnip greens. To keep the example small, we have used only the data for $n = 4$ determinations on each of $a = 4$ leaves. In the analysis of variance (shown below table 10.13.1), the mean square between leaves s_L^2 is an unbiased estimate of $\sigma^2 + n\sigma_A^2 = \sigma^2 + 4\sigma_A^2$. Consequently, an unbiased estimate of σ_A^2 is

$$s_A^2 = (s_L^2 - s^2)/4 = (0.2961 - 0.0066)/4 = 0.0724$$

The quantity σ_A^2 is called the *component of variance* for leaves. The value of $F = 0.2961/0.0066 = 44.9$ (highly significant with 3 and 12 *d.f.*) is an estimate of $(\sigma^2 + 4\sigma_A^2)/\sigma^2$.

We now consider the questions: (i) How precisely has the mean calcium content been estimated? (ii) Can we estimate it more economically? With n determinations from each of a leaves, the sample mean $\bar{X}_{..}$ is, from equation 10.13.1 for model II,

$$\bar{X}_{..} = \mu + \bar{A}_{.} + \bar{\varepsilon}_{..},$$

where $\bar{A}_{.}$ is the mean of a independent values of A_i (one for each leaf), and $\bar{\varepsilon}_{..}$ is the mean of an independent ε_{ij} . Hence the variance of $\bar{X}_{..}$ as an estimate of μ is

$$V(\bar{X}_{..}) = \frac{\sigma_A^2}{a} + \frac{\sigma^2}{an} = \frac{\sigma^2 + n\sigma_A^2}{an} = \frac{\sigma^2 + 4\sigma_A^2}{16} \quad (10.13.4)$$

TABLE 10.13.1
CALCIUM CONCENTRATION IN TURNIP GREENS
(Per cent of dry weight)

Leaf	Per Cent of Calcium				Sum	Mean
1	3.28	3.09	3.03	3.03	12.43	3.11
2	3.52	3.48	3.38	3.38	13.76	3.44
3	2.88	2.80	2.81	2.76	11.25	2.81
4	3.34	3.38	3.23	3.26	13.21	3.30

Source of Variation	Degrees of Freedom	Mean Square	Parameters Estimated
Between leaves	3	0.2961	$\sigma^2 + 4\sigma_A^2$
Determinations	12	0.0066	σ^2

$$s^2 = 0.0066 \text{ estimates } \sigma^2 \quad s_A^2 = (0.2961 - 0.0066)/4 = 0.0724 \text{ estimates } \sigma_A^2$$

In the analysis of variance, the mean square between leaves, 0.2961, is an unbiased estimate of $(\sigma^2 + 4\sigma_A^2)$. Hence, $\hat{V}(\bar{X}_{..}) = (0.2961)/16 = 0.0185$. This is an important result. The estimated variance of the sample mean is the *Between classes mean square, divided by the total number of observations*.

Suppose that the experiment is to be redesigned, changing n and a to n' and a' . As in equation 10.13.4, the variance of $\bar{X}_{..}$ becomes

$$V'(\bar{X}_{..}) = \frac{\sigma_A^2}{a'} + \frac{\sigma^2}{a'n'} \sim \frac{0.0724}{a'} + \frac{0.0066}{a'n'},$$

where the \sim sign means "is estimated by." Since the larger numerator is 0.0724, it seems clear that a' should be increased and n' decreased if this is possible without increasing the total cost of the experiment. If a determination costs 10 times as much as a leaf, the choice of $n' = 1$ and $a' = 15$ will cost about the same as our original data. For this new design our estimate of the variance of $\bar{X}_{..}$ is

$$\hat{V}'(\bar{X}_{..}) = \frac{0.0724}{15} + \frac{0.0066}{15} = 0.0053$$

The change reduces the variance of the mean from 0.0185 to 0.0053, i.e., to less than one-third. This is because the costly determinations with small variability have been utilized to sample more leaves whose variation is large. A formula for determining the best values of a' and n' in a given cost situation will be found in sections 17.11 and 17.12.

With model II, the difference $(X_{ij} - \mu)$ between a single observation and the population mean is the sum of the two terms A_i and ε_{ij} . Hence, the variance of X_{ij} is $(\sigma_A^2 + \sigma^2)$. The two parts are called the *components of variance*. The previous example illustrates how these components are used in problems of measurement, the objective being to estimate μ as

economically as possible. In plant breeding, n replications of each of a inbred lines may be grown in an experiment. The component σ_A^2 represents differences in yield that are due to differences in the genotypes (genetic characteristics) of the inbreds, while σ^2 measures the effect of non-genetic influences on yield. The ratio $\sigma_A^2/(\sigma_A^2 + \sigma^2)$ of genetic to total variance gives a guide to the possibility of improving yield by selection of particular inbreds. The same concepts are important in human family studies, both in genetics and the social sciences, where the ratio $\sigma_A^2/(\sigma_A^2 + \sigma^2)$ now measures the proportion of the total variance that is associated with the family. The interpretation is more complex, however, since human families differ not only in genetic traits but also in environmental factors that affect the variables under study.

EXAMPLE 10 13 1—The following data were abstracted from records of performance of Poland China swine in a single inbred line at the Iowa Agricultural Experiment Station. Two boars were taken from each of four litters with common sire and fed a standard ration from weaning to about 225 pounds. Here are the average daily gains

Litter	1	2	3	4
Gains	1 18 1 11	1 36 1 65	1 37 1 40	1 07 0 90

Assuming that the litter variable is normally distributed, show that σ_A^2 differs significantly from zero ($F = 7.41$) and that 0.0474 estimates it.

EXAMPLE 10 13 2—There is evidence that persons estimating the crop yields of fields by eye tend to underestimate high yields and overestimate low yields. If so, and if two estimators make separate estimates of the yields of each of a number of fields, what will be the effect on (i) the model II assumptions, (ii) the estimate s_A^2 of the variance σ_A^2 between fields, (iii) the estimate s^2 of σ^2 ?

EXAMPLE 10 13 3—To prove the result (10 13 3) for the expected value of the mean square between classes, show that under model II,

$$(\bar{X}_i - \bar{X}) = (A_i - \bar{A}) + (\varepsilon_i - \varepsilon)$$

$$\frac{\sum (\bar{X}_i - \bar{X})^2}{(a-1)} = \frac{\sum (A_i - \bar{A})^2}{(a-1)} + \frac{\sum (\varepsilon_i - \varepsilon)^2}{(a-1)} + \frac{2\sum (A_i - \bar{A})(\varepsilon_i - \varepsilon)}{(a-1)},$$

where \bar{X}_i is the mean of the n determinations in class i , and \bar{X} is the overall sample mean. If a random sample of leaves has been drawn, the first term on the right is an unbiased estimate of σ_A^2 , and the second of σ^2/n , since ε_i is the mean of n independent determinations.

The third term vanishes, on the average in repeated sampling, if the A_i and ε_i are independent. Multiplying by n to obtain the mean square between classes, the result follows. See if you can obtain the corresponding result (10 13 2) for model I.

10.14—Structure of model II illustrated by sampling. It is easy to construct a model II experiment by sampling from known populations. One population can be chosen to represent the individuals with variance σ^2 and another to represent the variable class effects with variance σ_A^2 , then samples can be drawn from each and combined in any desired

TABLE 10 14 1
GAINS IN WEIGHT OF 20 PIGS IN TEN LITTERS OF TWO PIGS EACH
(Each gain is the sum of three components The component for litters is a sample
with $\sigma_A^2 = 25$, that for individuals is from table 3 2 1 with $\sigma^2 = 100$)

Litter Number	Litter Component A_i	Pig Component ε_{ij}	Sample of Pig Gains $X_{ij} = \mu + A_i + \varepsilon_{ij}$	Sample of Litter Gains
(1)	(2)	(3)	(4) = 30 + (2) + (3)	(5)
1	- 1	7 9	36 38	74
2	2	- 4 -23	28 9	37
3	- 1	0 19	29 48	77
4	0	2 2	32 32	64
5	- 4	3 12	29 38	67
6	-10	9 3	29 23	52
7	10	5 - 4	45 36	81
8	2	-19 -10	13 22	35
9	4	- 4 18	30 52	82
10	- 2	15 - 6	43 22	65

Source of Variation	Degrees of Freedom	Mean Square	Parameters Estimated
Litters	9	144.6	$\sigma^2 + 2\sigma_A^2$
Individuals	10	96.5	σ^2

$$s^2 = 96.5 \text{ estimates } 100 \quad s_A^2 = (144.6 - 96.5)/2 = 24.0 \text{ estimates } 25$$

proportion In table 10 14 1 is such a drawing The sample consists of two pigs from each of ten litters the litters simulating random class effects Individual pig gains were taken from table 3 2 1 with $\sigma^2 = 100$, two of these per litter The litter components were drawn from a population with $\sigma_A^2 = 25$ (table 3 10 1 in the fifth edition of this book)

The usual analysis of variance is computed from table 10.14.1, then the components of variance are separated. From the 20 observations we obtained estimates $s^2 = 96.5$ of $\sigma^2 = 100$ and $s_A^2 = 24.0$ of $\sigma_A^2 = 25$, the two components that were put into the data.

This example was chosen because of its accurate estimates. An idea of ordinary variation can be got from examination of the records of 25 similar samples in table 10.14.2. One is struck immediately by the great variability in the estimates of σ_A^2 , some of them being negative! These latter merely indicate that the mean square for litters is less than that for individuals; the litters vary less than random samples ordinarily do if drawn from a single, normal population. Clearly, one cannot hope for accurate estimates of σ^2 and σ_A^2 from such small samples.

TABLE 10.14.2
ESTIMATES OF $\sigma_A^2 = 25$ AND $\sigma^2 = 100$ MADE FROM 25 SAMPLES DRAWN LIKE
THAT OF TABLE 10.14.1

Sample Number	Estimate of $\sigma_A^2 = 25$	Estimate of $\sigma^2 = 100$	Sample Number	Estimate of $\sigma_A^2 = 25$	Estimate of $\sigma^2 = 100$
1	60	127	14	56	112
2	56	104	15	-33	159
3	28	97	16	67	54
4	6	91	17	-18	90
5	18	60	18	33	65
6	-5	91	19	-21	127
7	7	53	20	-48	126
8	-1	87	21	4	43
9	0	66	22	3	145
10	-78	210	23	49	142
11	14	148	24	75	23
12	7	162	25	77	106
13	68	76			
			Mean	17.0	102.6

EXAMPLE 10.14.1—In table 10.14.2, how many negative estimates of σ_A^2 would be expected? Ans. A negative estimate occurs whenever the observed $F < 1$. From section 10.13, the probability that the observed $F < 1$ is the probability that the ordinary $F < 1/(1 + 2\sigma_A^2/\sigma^2)$, or in this example, $< 1/1.5 = 2/3$, where F has 9 and 10 df . A property of the F distribution is that this probability is the probability that F , with 10 and 9 df , exceeds $3/2$, or 1.5. From table A.14, with $f_1 = 10$, $f_2 = 9$, we see that F exceeds 1.59 with $P = 0.25$. Thus about $(0.25)(25) = 6.2$ negative estimates are expected, as against 7 found in table 10.14.2.

10.15—Confidence limits for σ_A^2 . Assuming normality, approximate confidence limits for σ_A^2 have been given by Morignuti (18). We shall illustrate from the turnip greens example (table 10.13.1) for which $n = 4$, $f_1 = 3$, $f_2 = 12$, $s_A^2 = 0.0724$, and $s^2 = 0.0066$. It is necessary to look up four entries in the F -table. If the table of 5% significance levels is used, these determine a two-tailed 90% confidence interval, with 5% on each tail. The 5% values of F needed are as follows:

$$\begin{aligned}
F_1 &= F_{f_1, f_2} = F_{3, 12} &= 3.49 \\
F_2 &= F_{f_1, \infty} = F_{3, \infty} &= 2.60 \\
F_3 &= F_{f_2, f_1} = F_{12, 3} &= 8.74 \\
F_4 &= F_{\infty, f_1} = F_{\infty, 3} &= 8.53 \\
F &= \text{observed value of } F = 44.9
\end{aligned}$$

The limits are given as multipliers of the quantity $s^2/n = (0.0066)/4 = 0.00165$. The lower limit for σ_A^2 is

$$\begin{aligned}
\hat{\sigma}_{AL}^2 &= \frac{(F - F_1)(F + F_1 - F_2)}{FF_2} \frac{s^2}{n} = \frac{(44.9 - 3.49)(44.9 + 3.49 - 2.60)}{(44.9)(2.60)} (0.00165) \\
&= \frac{(41.41)(45.79)}{(44.9)(2.60)} (0.00165) = 0.027
\end{aligned}$$

As would be expected, the lower limit becomes zero if $F = F_1$; that is, if F is just significant at the 5% level.

The upper limit is

$$\begin{aligned}
\hat{\sigma}_{AU}^2 &= \left\{ FF_4 - 1 + \frac{(F_3 - F_4)}{FF_3} \right\} \frac{s^2}{n} \\
&= \{(44.9)(8.53) - 1 + (0.21)/(44.9)(8.74)\} (0.00165) = 0.63
\end{aligned}$$

Frequently, as in this example, the rather unwieldy second term inside the curly bracket is negligible and need not be computed.

To summarize, the estimate is $s_A^2 = 0.0724$, with 90% confidence limits 0.027 and 0.63. Earlier, Bross (19) gave approximate fiducial limits, using the same five values of F . His limits agree closely with the above limits whenever F is significant.

If the distributions of A_i and ε_{ij} are non-normal, having positive kurtosis, the variance of s_A^2 is increased, and the above confidence intervals are too narrow.

EXAMPLE 10.15.1—In estimating the amount of plankton in an area of sea, seven runs (called hauls) were made, with six nets on each run (20). Estimate the component of variance between hauls and its 90% confidence limits.

	Degrees of Freedom	Mean Square
Between hauls	6	0.1011
Within hauls	35	0.0208

Ans. $s_A^2 = 0.0134$, with limits (0.0044, 0.058).

10.16—Samples within samples. Nested classifications. Each sample may be composed of *sub-samples* and these in turn may be sub-sampled, etc. The repeated sampling and sub-sampling gives rise to *nested* or *hierarchal classifications*, as they are sometimes called.

In table 10.16.1 is an example. This is a part of the turnip greens experiment cited earlier (17). The four plants were taken at random, then three leaves were randomly selected from each plant. From each leaf were taken two samples of 100 mg. in which calcium was determined by microchemical methods. The immediate objective is to separate the sums of squares due to the sources of variation, plants, leaves of the same plant, and determinations on the leaves.

The calculations are given under table 10.16.1. The total sums of squares for determinations, leaves, and plants are first obtained by the usual formulas. The sum of squares *between leaves of the same plant* is found by subtracting the sum of squares between plants from that between leaves, as shown. Similarly, the sum of squares *between determina-*

TABLE 10.16.1
CALCIUM CONCENTRATION (PER CENT, DRY BASIS) IN $b = 3$ LEAVES FROM EACH OF
 $a = 4$ TURNIP PLANTS, $n = 2$ DETERMINATIONS PER LEAF. ANALYSIS OF VARIANCE

Plant, i $i = 1 \dots a$	Leaf, j $j = 1 \dots b$	Determinations, X_{ijk}		$X_{ij\cdot}$	$X_{i\cdot\cdot}$	$X_{\cdot\cdot\cdot}$
1	1	3.28	3.09	6.37	19.05	
	2	3.52	3.48	7.00		
	3	2.88	2.80	5.68		
2	1	2.46	2.44	4.90	13.07	
	2	1.87	1.92	3.79		
	3	2.19	2.19	4.38		
3	1	2.77	2.66	5.43	17.71	
	2	3.74	3.44	7.18		
	3	2.55	2.55	5.10		
4	1	3.78	3.87	7.65	22.46	72.29
	2	4.07	4.12	8.19		
	3	3.31	3.31	6.62		

Total Size = $abn = (4)(3)(2) = 24$ determinations

$$C = (X_{\cdot\cdot\cdot})^2 / abn = (72.29)^2 / 24 = 217.7435$$

$$\text{Determinations } \Sigma X_{ijk}^2 - C = 3.28^2 + \dots + 3.31^2 - C = 10.2704$$

$$\text{Leaves } \Sigma X_{ij\cdot}^2 / n - C = (6.37^2 + \dots + 6.62^2) / 2 - C = 10.1905$$

$$\text{Plants } \Sigma X_{i\cdot\cdot}^2 / b - C = (19.05^2 + \dots + 22.46^2) / 4 - C = 7.5603$$

$$\text{Leaves of the same plant} = \text{Leaves} - \text{Plants} = 10.1905 - 7.5603 = 2.6302$$

$$\text{Determinations on same leaf} = \text{Determinations} - \text{Leaves} = 10.2704 - 10.1905 = 0.0799$$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Plants	3	7.5603	2.5201
Leaves in plants	8	2.6302	0.3288
Determinations in leaves	12	0.0799	0.0067
Total	23	10.2704	

tions on the same leaf is obtained by deducting the total sum of squares between leaves from that between determinations. This process can be repeated with successive sub-sampling.

The model being used is,

$$X_{ijk} = \mu + A_i + B_{ij} + \varepsilon_{ijk}, i = 1 \dots a, j = 1 \dots b, k = 1 \dots n, \\ A_i = \mathcal{N}(0, \sigma_A), B_{ij} = \mathcal{N}(0, \sigma_B), \varepsilon_{ijk} = \mathcal{N}(0, \sigma), \quad (10.16.1)$$

where A refers to plants and B to leaves. The variables A_i , B_{ij} , and ε_{ijk} are all assumed independent. Roman letters are used to denote plants and leaves because they are random variables, not constants.

TABLE 10.16 2
COMPLETED ANALYSIS OF VARIANCE OF TURNIP GREENS DATA

Source of Variation	Degrees of Freedom	Mean Square	Parameters Estimated
Plants	3	2 5201	$\sigma^2 + n\sigma_B^2 + bn\sigma_A^2$
Leaves in plants	8	0 3288	$\sigma^2 + n\sigma_B^2$
Determinations in leaves	12	0.0067	σ^2

$n = 2, b = 3, a = 4$ $s^2 = 0.0067$ estimates σ^2 . $s_B^2 = (0.3288 - 0.0067)/2 = 0.1610$ estimates σ_B^2 $s_A^2 = (2\ 5201 - 0\ 3288)/6 = 0\ 3652$ estimates σ_A^2 .

In the completed analysis of variance, table 10.16.2, the components of variance are shown. Each component in a sub-sample is included among those in the sample above it. The estimates are calculated as indicated.

Null hypotheses which may be tested are:

- $\sigma_A^2 = 0$; $F = \frac{2.5201}{0.3288} = 7.66$ estimates $\frac{\sigma^2 + n\sigma_B^2 + bn\sigma_A^2}{\sigma^2 + n\sigma_B^2}$, $f = 3, 8$.
- $\sigma_B^2 = 0$; $F = \frac{0.3288}{0.0067} = 49$ estimates $\frac{\sigma^2 + n\sigma_B^2}{\sigma^2}$, $f = 8, 12$.

For the first, with degrees of freedom, $f_1 = 3$ and $f_2 = 8$, F is almost on its 1% point, 7.59; for the second, with degrees of freedom 8 and 12, F is far beyond its 1% point, 4.50. Evidently, in the sampled population the per cent calcium varies both from leaf to leaf and from plant to plant.

As with a single sub-classification (plants and leaves in section 10.13), it may be shown that the estimated variance of the sample mean per determination is given by the mean square *between plants*, divided by the number of determinations. This estimated variance can be expressed in terms of the estimated components of variance from table 10.16.2, as follows:

$$s_{\bar{x}}^2 = \frac{2.5201}{24} = 0.105 = \frac{0.0067 + n(0.1610) + bn(0.3652)}{nab} \\ = \frac{0.0067}{nab} + \frac{0.1610}{ab} + \frac{0.3652}{a}$$

This suggests that more information per dollar may be got by decreasing n , the number of expensive determinations per leaf which have a small component, then increasing b or a , the numbers of leaves or plants. Plants presumably cost more than leaves, but the component is also larger. How to balance these elements is the topic of section 17.12.

Confidence limits for σ_A^2 and σ_B^2 are calculated by the method described in section 10.15.

EXAMPLE 10.16.1—Verify that the sum of squares for Determinations in leaves, as found by subtraction in table 10.16.1, is the sum of squares of deviations of the determinations from their respective leaf means: Ans. Since the C term cancels, Determinations — Leaves is equal to

$$\sum_i \sum_j \sum_k X_{ijk}^2 - \sum_i \sum_j X_{ij.}^2 / n = \sum_i \sum_j \sum_k (X_{ijk} - \bar{X}_{ij.})^2$$

by the usual shortcut rule for finding a sum of squares of deviations, where $\bar{X}_{ij.}$ is the mean of the n determinations on the j th leaf of the i th plant.

EXAMPLE 10.16.2—From equation 10.16.1 for the model, show that the variance of the sample mean is $(\sigma^2 + n\sigma_B^2 + bn\sigma_A^2)/abn$, and that an unbiased estimate of it is given by the mean square between plants, divided by abn , i.e., by $2.5201/24 = 0.105$, as stated in section 10.16.

EXAMPLE 10.16.3—If one determination were made on each of two leaves from each of ten plants, what is your estimate of the variance of the sample mean? Ans. 0.045.

EXAMPLE 10.16.4—With one determination on one leaf from each plant, how many plants must be taken in order to reduce $s_{\bar{X}}$ to 0.2? Ans. About 14. (This estimate is very rough, since the mean square between plants has only 3 d.f.)

10.17—Samples within samples. Mixed model. In some applications of sub-sampling, the major classes have fixed effects that are to be estimated. An instance is an evaluation of the breeding value of a set of five sires in pig-raising. Each sire is mated to a random group of dams, each mating producing a litter of pigs whose characteristics are the criterion. The model is:

$$X_{ijk} = \mu + \alpha_i + B_{ij} + \varepsilon_{ijk} \quad (10.17.1)$$

The α_i are constants ($\sum \alpha_i = 0$) associated with the sires but the B_{ij} and the ε_{ijk} are random variables corresponding to dams and offspring. Hence the model is called *mixed*.

Table 10.17.1 is an example with $b = 2$ dams for each sire and $n = 2$ pigs chosen from each litter for easy analysis (from records of the Iowa Agricultural Experiment Station). The calculations proceed exactly as in the preceding section. The only change is that in the mean square for sires, the term $nb\kappa^2$, where $\kappa^2 = \Sigma \alpha^2 / (a - 1)$, replaces $nb\sigma_A^2$.

In a mixed model of this type, two points must be noted. From equation 10.17.1, the observed class mean may be written

$$\bar{X}_{i..} = \mu + \alpha_i + \bar{B}_{i.} + \bar{\varepsilon}_{i..}$$

where $\bar{B}_{i.}$ is the average of b values of the B_{ij} and $\bar{\varepsilon}_{i..}$ is the average of nb values of the ε_{ijk} . Thus the variance of $\bar{X}_{i..}$ considered as an estimate of $\mu + \alpha_i$, is

TABLE 10.17.1
AVERAGE DAILY GAIN OF TWO PIGS OF EACH LITTER

Sire	Dam	Pig Gains		Sums		
1	1	2.77	2.38	5.15	10.67	
	2	2.58	2.94	5.52		
2	1	2.28	2.22	4.50	10.12	
	2	3.01	2.61	5.62		
3	1	2.36	2.71	5.07	10.53	
	2	2.72	2.74	5.46		
4	1	2.87	2.46	5.33	9.88	
	2	2.31	2.24	4.55		
5	1	2.74	2.56	5.30	10.28	51.48
	2	2.50	2.48	4.98		
Source of Variation		Degrees of Freedom		Mean Square	Parameters Estimated	
Sires		4		0.0249	$\sigma^2 + n\sigma_B^2 + nb\kappa^2$	
Dams-Same Sire		5		0.1127	$\sigma^2 + n\sigma_B^2$	
Pairs-Same Dam		10		0.0387	σ^2	
$n = 2, b = 2, s^2 = 0.0387$ estimates $\sigma^2, s_B^2 = (0.1127 - 0.0387)/2 = 0.0370$ estimates $\sigma_B^2, 0$ estimates κ^2						
To test $\sigma_B^2 = 0, F = 0.1127/0.0387 = 2.91, F_{0.05} = 3.33.$						

$$V(\bar{X}_{i..}) = \frac{\sigma_B^2}{b} + \frac{\sigma^2}{nb} = \frac{1}{nb}(\sigma^2 + n\sigma_B^2)$$

The analysis of variance shows that the mean square *between dams of the same sire* is the relevant mean square, being an unbiased estimate of $(\sigma^2 + n\sigma_B^2)$. The standard error of a sire mean is $\sqrt{(0.1127/4)} = 0.168$, with 5 *df*. Secondly, the *F* ratio for testing the null hypothesis that all α_i are zero is the ratio $0.0249/0.1127$. Since this ratio is substantially less than 1, there is no indication of differences between sires in these data.

10.18—Samples of unequal sizes. Random effects. This case occurs commonly in family studies in human and animal genetics and in the social sciences. The model being used is a form of model II:

$$X_{ij} = \mu + A_i + \varepsilon_{ij}, \quad i = 1, \dots, a, j = 1, \dots, n_i, \quad A_i = \mathcal{N}(0, \sigma_A), \quad \varepsilon_{ij} = \mathcal{N}(0, \sigma)$$

The new feature is that n_i , the size of sample of the *i*th class, varies from class to class. The total sample size is $N = \sum n_i$. All A_i and ε_{ij} are assumed independent.

The computations for the analysis of variance and the *F*-test of the null hypothesis $\sigma_A = 0$ are the same as for fixed effects, as given in section

10.12. With equal $n_i (=n)$, the mean square between classes was found to be an unbiased estimate of $\sigma^2 + n\sigma_A^2$ (section 10.13). With unequal n_i , the corresponding expression is $\sigma^2 + n_0\sigma_A^2$, where

$$n_0 = \frac{1}{(a-1)} \left(N - \frac{\sum n_i^2}{N} \right) = \bar{n} - \sum (n_i - \bar{n})^2 / (a-1)N$$

The first equation is the form used for computing n_0 . The second equation shows that n_0 is always *less* than the arithmetic mean \bar{n} of the n_i , although usually only slightly less.

Consequently, if s_b^2 and s^2 are the mean squares between and within classes, respectively, unbiased estimates of the two components of variance σ^2 and σ_A^2 are given by

$$\hat{\sigma}^2 = s^2 \quad : \quad \hat{\sigma}_A^2 = (s_b^2 - s^2)/n_0$$

With unequal n_i , some mathematical complexities arise that have not yet been overcome in a form suitable for practical use. The estimate $\hat{\sigma}_A^2$, while unbiased whether the A_i and ε_{ij} are normally distributed or not, is not fully efficient unless σ_A^2 is small. The method given for finding confidence limits for σ_A^2 with equal n (section 10.15) does not apply. An ingenious method of finding confidence limits for the ratio σ_A^2/σ^2 was, however, given by Wald (21). Whenever feasible, it pays to keep the sample sizes equal.

EXAMPLE 10.18.1—In research on artificial insemination of cows, a series of semen samples from a bull are sent out and tested for their ability to produce conceptions. The following data from a larger set kindly supplied by Dr. G. W. Salisbury, show the percentages of conceptions obtained from the samples for six bulls. In the analysis of variance, the total sum of squares, uncorrected, was 111,076. Verify the analysis of variance, the value of n_0 , and the estimates of the two variance components. (Since the data are percentages based on slightly differing numbers of tests, the assumption that σ^2 is constant in these data is not quite correct.)

Bull (i)	Percentages of Conceptions to Services for Successive Samples	n_i	X_i
1	46, 31, 37, 62, 30	5	206
2	70, 59	2	129
3	52, 44, 57, 40, 67, 64, 70	7	394
4	47, 21, 70, 46, 14	5	198
5	42, 64, 50, 69, 77, 81, 87	7	470
6	35, 68, 59, 38, 57, 76, 57, 29, 60	9	479
Total		35	1876

Source	df	$S.S.$	$M.S.$	$E(M.S.)$
Between bulls	5	3,772	754	$\sigma^2 + 5.67\sigma_A^2$
Within bulls	29	6,750	233	σ^2

$$s^2 = 233 \text{ estimates } \sigma^2 \quad (754 - 233)/5.67 = 92 \text{ estimates } \sigma_A^2$$

EXAMPLE 10.18.2—The preceding example is one in which we might consider either fixed or random effects of bulls, depending on the objectives. If these six bulls were available for an artificial insemination program, we would be interested in comparing the percentages of success of these specific bulls in a fixed effects analysis.

10.19—Samples within samples. Unequal sizes. Both samples and sub-samples may be of unequal sizes. Computational methods for any number of levels (samples, sub-samples, sub-sub-samples, etc.) have been developed by Gower (22) and by Gates and Shine (23), following earlier work by Ganguli (24): The analysis of variance is straightforward although tedious. A general procedure for finding unbiased estimates of the components of variance at each level will be given.

Our example is from a small survey of wheat yields in six districts in England (25). One or more farms were selected in each district, and from one to three wheat fields from each selected farm. Strictly, this is a mixed model, since the districts are fixed; further, the farms within districts were not randomly selected. The data serve, however, to illustrate the computations.

The computations are most easily followed if the data are set out as in table 10.19.1. The lowest level (fields) is denoted by 0. The yield, X_{0k} , and the number of observations in each yield are written down. In this example, as in most applications, the N_{0k} are all 1, each observation being the yield of one field.

The X_{0k} and the N_{0k} are added to give the totals, X_{1k} and N_{1k} , at the next lowest level, farms. Similarly, the X_{1k} and the N_{1k} are added to give the district totals, X_{2k} and N_{2k} . Finally, the district totals are added to give X_{3k} and N_{3k} , the grand total and the total number of recorded observations, respectively.

To obtain the sum of squares in the analysis of variance, first calculate for each level the quantity

$$S_i = \sum_k X_{ik}^2 / N_{ik}$$

S_3 , for instance, is $(1063)^2/36 = 31,388.0$, the usual correction for the mean. At Level 2 (Districts) we have

$$S_2 = 110^2/4 + 91^2/3 + \dots + 432^2/13 = 31,849.3$$

To obtain the *d.f.*, count the number of classes C_i at each level. These are $C_0 = 36$, $C_1 = 25$, $C_2 = 6$, $C_3 = 1$, as shown at the foot of table 10.19.1. The C_i and the S_i provide the *d.f.* and the sums of squares in the analysis of variance, as shown in table 10.19.2 on p. 293.

The rule for calculating the *d.f.* and the sums of squares is a straightforward extension of the rule for two levels given in table 10.12.1.

We now express the expected values of the three mean squares in terms of the components of variance for districts (σ_2^2), farms (σ_1^2), and fields (σ_0^2). For this we use two sets of auxiliary quantities, γ_{ij} and k_{ij} .

TABLE 10.19.1
WHEAT YIELDS (GMS. PER 0.0000904 ACRE) TO ILLUSTRATE ESTIMATION OF COMPONENTS
OF VARIANCE IN NESTED CLASSIFICATIONS WITH UNEQUAL NUMBERS

Level 0 Fields		Level 1 Farms		Level 2 Districts		Level 3 Grand Total	
X_{0k}	N_{0k}	X_{1k}	N_{1k}	X_{2k}	N_{2k}	X_{3k}	N_{3k}
23	1						
19	1	42	2				
31	1						
37	1	68	2	110	4		
33	1						
29	1	62	2				
29	1	29	1	91	3		
36	1						
29	1						
33	1	98	3	98	3		
11	1						
21	1	32	2				
23	1						
18	1	41	2				
33	1	33	1				
23	1	23	1				
26	1	26	1				
39	1	39	1				
20	1	20	1				
24	1	24	1				
36	1	36	1	274	11		
25	1						
33	1	58	2	58	2		
28	1						
31	1	59	2				
25	1						
42	1	67	2				
32	1						
36	1	68	2				
41	1	41	1				
35	1	35	1				
16	1	16	1				
30	1	30	1				
40	1	40	1				
32	1	32	1				
44	1	44	1	432	13	1063	36
C_i	36		25		6		1

For the γ_{ij} , i and j take the values 0, 1, 2, 3, with $i \geq j$. In the diagonal, γ_{ii} always equals the total number of observations, in this case 36. Further, when all N_{0k} are 1, $\gamma_{i0} = C_i$, the number of classes at level i . Thus, we write 1, 6, 25, and 36 in the column γ_{i0} in table 10.19.3. For the remaining γ_{ij} , the rule is (using table 10.19.1):

Sum the squares of the N_{jk} , each square divided by the next entry N_{ik} at level i . It sounds puzzling but should be clear from the examples.

TABLE 10.19.2
ANALYSIS OF VARIANCE OF WHEAT YIELDS

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares
Districts (level 2)	$C_2 - C_3 = 5$	$S_2 - S_3 = 461.3$	92.3
Farms within districts (level 1)	$C_1 - C_2 = 19$	$S_1 - S_2 = 1,349.5$	71.0
Fields within farms (level 0)	$C_0 - C_1 = 11$	$S_0 - S_1 = 310.2$	28.2

TABLE 10.19.3
VALUES OF AUXILIARY QUANTITIES γ_{ij} AND k_{ij}

$i \backslash j$	0	1	2	3
3	1	1.67	9.11	36
2	6	11.49	36	
1	25	36		
0	36			

$i \backslash j$	0	1	2
2	5	9.82	26.89
1	19	24.51	
0	11		

$$\gamma_{21} = \frac{2^2 + 2^2}{4} + \frac{2^2 + 1^2}{3} + \frac{3^2}{3} + \dots + \frac{2^2 + 2^2 + \dots + 1^2 + 1^2}{13} = 11.49$$

$$\gamma_{32} = (4^2 + 3^2 + 3^2 + 11^2 + 2^2 + 13^2)/36 = 9.11$$

For the k_{ij} , i and j take the values 0, 1, 2, with $i \geq j$, and

$$k_{ij} = \gamma_{ij} - \gamma_{i+1,j}$$

That is, to find any k_{ij} , start with γ_{ij} and subtract the number immediately above it. Thus, $k_{22} = 36 - 9.11 = 26.89$.

The quantity k_{ij} is the coefficient of σ_j^2 in the expected value of the sum of squares at level i in the analysis of variance. To find the expected values of the corresponding mean squares, divide by the number of df . at level i . These mean squares (from table 10.19.2) and their expected values appear in table 10.9.4. For example, the coefficient 1.290 of σ_1^2 in the farms mean square is $k_{11}/19 = 24.51/19$, and so on.

TABLE 10.19.4
EXPECTED VALUES OF THE MEAN SQUARES

Level	Degrees of Freedom	Mean Square	Expected Value
Districts ($i = 2$)	5	92.3	$\sigma_0^2 + 1.964\sigma_1^2 + 5.378\sigma_2^2$
Farms ($i = 1$)	19	71.0	$\sigma_0^2 + 1.290\sigma_1^2$
Fields ($i = 0$)	11	28.2	σ_0^2

A new feature is that the coefficient of σ_1^2 is no longer the same in the Districts and Farms mean squares. Thus, the ratio $92.3/71.0$ cannot be used as an F -test of the null hypothesis $\sigma_2^2 = 0$. However, unbiased

estimates of the three components are obtained from table 10.19.4 as follows:

$$\begin{aligned}s_0^2 &= 28.2 & : & \quad s_1^2 = (71.0 - 28.2)/1.290 = 33.2 \\ s_2^2 &= [92.3 - 28.2 - (1.964)(33.2)]/5.378 = -0.02\end{aligned}$$

The data give no evidence of real differences in yield between districts.

This method of calculation holds for any number of levels. For large bodies of data the computations may be programmed for an electronic computer.

10.20—Intraclass correlation. We revert to a single classification with n members per class. When the component $\sigma_A^2 > 0$, we have seen that members of the same class tend to act alike. An alternative to model II for describing this situation is to suppose that the observations X_{ij} are all distributed about the same mean μ with the same variance σ^2 , but that any two members of the same class ($i = \text{constant}$) have a common correlation coefficient ρ_I , called the *intraclass correlation coefficient*. Actually, this model antedates the analysis of variance.

With this model it can be shown by algebra that the expected values of the mean squares in the analysis of variance are as follows:

Source of Variation	Mean Square	Expected Value
Between classes	s_b^2	$\sigma^2\{1 + (n-1)\rho_I\}$
Within classes	s_w^2	$\sigma^2(1 - \rho_I)$

This model is useful in applications in which it is natural to think of members of the same class as correlated. It is frequently employed in studies of twins ($n = 2$). The model is more general than the components of variance model. If ρ_I is negative, note that s_b^2 has a *smaller* expected value than s_w^2 . With model II, this cannot happen. But if, for instance, four young animals in a pen compete for an insufficient supply of food, the stronger animals may drive away the weaker and may regularly get most of the food. For this reason the variance in weight within pens may be larger than that between pens, this being a real phenomenon and not an accident of sampling. We say that there is a negative correlation ρ_I between the weights within a pen. One restriction on negative values of ρ_I is that ρ_I cannot be less than $-1/(n-1)$. This is so because the expected value of s_b^2 must be greater than or equal to zero.

From the analysis of variance it is clear that $(s_b^2 - s_w^2)$ estimates $n\rho_I\sigma^2$, while $\{s_b^2 + (n-1)s_w^2\}$ estimates $n\sigma^2$. This suggests that as an estimate of ρ_I we take

$$r_I = (s_b^2 - s_w^2)/\{s_b^2 + (n-1)s_w^2\} \quad (10.20.1)$$

As will be seen presently, a slightly different estimate of ρ_I is obtained when we approach the problem from the viewpoint of correlation.

The data on identical twins in table 10.20.1 illustrate a high positive

TABLE 10.20.1
NUMBER OF FINGER RIDGES ON BOTH HANDS OF INDIVIDUALS IN 12 PAIRS
OF FEMALE IDENTICAL TWINS
[Data from Newman, Freeman, and Holzinger (34)]

Pair	Finger Ridges of Individuals	Pair	Finger Ridges of Individuals	Pair	Finger Ridges of Individuals
1	71, 71	5	76, 70	9	114, 113
2	79, 82	6	83, 82	10	94, 91
3	105, 99	7	114, 113	11	75, 83
4	115, 114	8	57, 44	12	76, 72

Analysis of Variance

Source of Variation	Degrees of Freedom	Mean Square
Twinn pairs	11	817.31
Individuals	12	14.29

$$s^2 = 14.29, \quad s_A^2 = 401.51, \quad r_I = 0.966$$

correlation. The numbers of finger ridges are nearly the same for the two members of each pair but differ markedly among pairs. From the analysis of variance, the estimate of ρ_I is ($n = 2$)

$$r_I = (817.31 - 14.29)/(817.31 + 14.29) = 0.966$$

In chapter 7, the ordinary correlation coefficient between X and Y was estimated as

$$r = \Sigma(X - \bar{X})(Y - \bar{Y}) / \sqrt{\{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2\}}$$

With twin data, which member of a pair shall we call X and which Y ? The solution is to count each point *twice*, once with the first member of a pair as X , and once with the first member as Y . Thus, pair 2 is entered as (79, 82) and also as (82, 79), while pair 1, where the order makes no difference, is entered as (71, 71) twice. With this method the X and Y samples both have the same mean and the same variance. If (X, X') denote the observations for a typical pair, you may verify that the correlation coefficient becomes

$$r_I' = 2\Sigma(X - \bar{X})(X' - \bar{X}) / \{\Sigma(X - \bar{X})^2 + \Sigma(X' - \bar{X})^2\}$$

where the sums are over the a pairs and \bar{X} is the mean of all observations. For the finger ridges, $r_I' = 0.962$.

With pairs ($n = 2$), intraclass correlations may be averaged and may have confidence limits set by using the transformation from r to z in section 7.7. The only changes are: (i) the variance of z_I is $1/(a - 3/2)$, where a is the number of pairs, as against $1/(a - 3)$ with an ordinary z , (ii) the correction for the bias in z_I is to add $1/(2a - 1)$.

With triplets ($n = 3$), each trio X, X', X'' specifies six points: (X, X') , (X', X) , (X, X'') , (X'', X) , (X', X'') , (X'', X') . The number of points rises

rapidly as n rises, and this method of calculating r_I' becomes discouraging. In 1913, however, Harris (26) discovered a shortened process similar to the analysis of variance, by showing in effect that

$$r_I' = \frac{(a-1)s_b^2 - as_w^2}{(a-1)s_b^2 + a(n-1)s_w^2}$$

Comparison with equation 10.20.1 shows that r_I' differs slightly from r_I , the difference being trivial unless a (the number of classes) is small. Since it is slightly simpler, equation (10.20.1) is more commonly used now as the sample estimate of ρ_I .

10.21—Tests of homogeneity of variance. From time to time we have raised the question as to whether two or more mean squares differ significantly. For two mean squares an answer, using the two-tailed F -test, was given in section 4.15. With more than two independent estimates of variance, Bartlett (27) provided a test.

If there are a estimates s_i^2 , each with the same number of degrees of freedom f , the test criterion is

$$M = 2.3026f(a \log \bar{s}^2 - \sum \log s_i^2) \quad (\bar{s}^2 = \sum s_i^2/a)$$

The factor 2.3026 is a constant ($\log_e 10$). On the null hypothesis that each s_i^2 is an estimate of the same σ^2 , the quantity M/C is distributed approximately as χ^2 with $(a-1)f$ d.f., where

$$C = 1 + \frac{a+1}{3af}$$

Since C is always slightly greater than 1, it need be used only if M lies close to one of the critical values of χ^2 .

In table 10.21.1 this test is applied to the variances of grams of fat absorbed in the four types of fat in the doughnut example of table 10.2.1. Here $a = 4$ and $f = 5$. The value of M is 1.88, clearly not significant with 3 d.f. To illustrate the method, $\chi^2 = M/C = 1.74$ has also been computed.

When the degrees of freedom differ, as with samples of unequal sizes, the computation of χ^2 is more tedious though it follows the same pattern. The formulas are:

$$M = (2.3026)[(\sum f_i) \log \bar{s}^2 - \sum f_i \log s_i^2] \quad (\bar{s}^2 = \sum f_i s_i^2 / \sum f_i)$$

$$C = 1 + \frac{1}{3(a-1)} \left[\sum \frac{1}{f_i} - \frac{1}{\sum f_i} \right]$$

$\chi^2 = M/C$ with $(a-1)$ degrees of freedom

In table 10.21.2 this test is applied to the variances of the birth weights of five litters of pigs. Since \bar{s}^2 is the pooled variance (weighting by degrees of freedom), we need a column of the sums of squares. A column of the reciprocals $1/f_i$ of the degrees of freedom is also useful in finding C . The

TABLE 10.21.1
COMPUTATION OF BARTLETT'S TEST OF HOMOGENEITY OF VARIANCE
ALL ESTIMATES HAVING $f = 5$ DEGREES OF FREEDOM

Fat	s_i^2	$\log s_i^2$
1	178	2.2504
2	60	1.7781
3	98	1.9912
4	68	1.8325
Total	404 $\bar{s}^2 = 100.9$	7.8522 $\log \bar{s}^2 = 2.0038$

$$M = (2.3026)(5)[4(2.0038) - 7.8522] = 1.88, (d.f. = 3)$$

$$C = 1 + \frac{a+1}{3af} = 1 + \frac{5}{(3)(4)(5)} = 1.083$$

$$\chi^2 = 1.88/1.083 = 1.74 (d.f. = 3), P > 0.5$$

computations give $\chi^2 = 16.99$ with 4 $d.f.$, showing that the intralitter variances differ from litter to litter in these data.

When some or all of the s_i^2 are less than 1, as in these data, it is worth noting that χ^2 is unchanged if all s_i^2 and \bar{s}^2 are multiplied by the same number (say 10 or 100). This enables you to avoid logs that are negative

TABLE 10.21.2
COMPUTATION OF BARTLETT'S TEST OF HOMOGENEITY OF VARIANCE.
SAMPLES DIFFERING IN SIZE

Litter (Sample)	Sum of Squares $f_i s_i^2$	Degrees of Freedom f_i	Mean Squares s_i^2	$\log s_i^2$	$f_i \log s_i^2$	Reciprocals $1/f_i$
1	8.18	9	0.909	-0.0414	-0.3726	0.1111
2	3.48	7	0.497	-0.3036	-2.1252	0.1429
3	0.68	9	0.076	-1.1192	-10.0728	0.1111
4	0.72	7	0.103	-0.9872	-6.9104	0.1429
5	0.73	5	0.146	-0.8357	-4.1785	0.2000
$a = 5$	13.79	37			-23.6595	0.7080

$$\bar{s}^2 = \Sigma f_i s_i^2 / \Sigma f_i = 13.79/37 = 0.3727$$

$$(\Sigma f_i) \log \bar{s}^2 = (37)(-0.4286) = -15.8582$$

$$M = (2.3026)[(\Sigma f_i) \log \bar{s}^2 - \Sigma f_i \log s_i^2]$$

$$= (2.3026)[-15.8582 - (-23.6595)] = 17.96$$

$$C = 1 + \frac{1}{(3)(4)} \left[0.7080 - \frac{1}{37} \right] = 1.057$$

$$\chi^2 = M/C = 17.96/1.057 = 16.99, (d.f. = 4) P < 0.01$$

The χ^2 approximation becomes less satisfactory if most of the f_i are less than 5. Special tables for this case are given in (28). This reference also gives a table of the significance levels of s_{\max}^2/s_{\min}^2 , the ratio of the largest to the smallest of the a variances. This ratio provides a quick test of homogeneity of variance which, though less sensitive than Bartlett's test, will often settle the issue.

Unfortunately, both Bartlett's test and this test are sensitive to non-normality in the data, particularly to kurtosis (29). With long-tailed distributions (positive kurtosis) the test gives too many erroneous verdicts of heterogeneity.

REFERENCES

1. B. LOWE. Data from the Iowa Agricultural Experiment Station (1935).
2. R. RICHARDSON, *et al.*. *J. Nutrition*, 44:371 (1951).
3. G. W. SNEDECOR. *Analysis of Variance and Covariance*. Collegiate Press, Inc., Ames, Iowa (1934).
4. R. A. FISHER and F. YATES. *Statistical Tables*. Oliver and Boyd, Edinburgh (1938).
5. T. R. HANSBERRY and C. H. RICHARDSON. *Iowa State Coll. J. Sci.*, 10:27 (1935).
6. ROTHAMSTED EXPERIMENTAL STATION REPORT: p. 289 (1936).
7. ROTHAMSTED EXPERIMENTAL STATION REPORT: p. 212 (1937).
8. R. A. FISHER. *The Design of Experiments*. Oliver and Boyd, Edinburgh (1935).
9. Query in *Biometrics*, 5:250 (1949).
10. D. NEWMAN. *Biometrika*, 31:20 (1939).
11. H. SCHEFFE. *The Analysis of Variance*. Wiley, New York (1959).
12. D. B. DUNCAN. *Ann. Math. Statist.*, 32:1013 (1961).
13. T. E. KURTZ, B. F. LINK, J. W. TUKEY, and D. L. WALLACE. *Technometrics*, 7:95 (1965).
14. G. E. P. BOX. *Ann. Math. Statist.*, 25:290 (1954).
15. P. C. TANG. *Statist. Res. Memoirs*, 2:126 (1938).
16. E. S. PEARSON and H. O. HARTLEY. *Biometrika*, 38:112 (1951).
17. "Studies of Sampling Techniques and Chemical Analyses of Vegetables." Southern Coop. Ser. Bull. 10 (1951).
18. S. MORIGUTI. *Reports of Statistical Applications in Research*, Japanese Union of Scientists and Engineers, Vol. 3, No. 2:29 (1954).
19. I. D. J. BROSS. *Biometrics*, 6:136 (1950).
20. C. P. WINSOR and G. L. CLARKE. *J. Marine Res.*, 3:1 (1940).
21. A. WALD. *Ann. Math. Statist.*, 11:96 (1940).
22. J. C. GOWER. *Biometrics*, 18:537 (1962).
23. C. E. GATES and C. SHINE. *Biometrics*, 18:529 (1962).
24. M. GANGULI. *Sankhyā*, 5:449 (1941).
25. W. G. COCHRAN. *Jour. Amer. Statist. Ass.*, 34:492 (1939).
26. J. A. HARRIS. *Biometrika*, 9:446 (1913).
27. M. S. BARTLETT. *Jour. Royal Statist. Soc. Suppl.*, 4:137 (1937).
28. E. S. PEARSON and H. O. HARTLEY. *Biometrika Tables for Statisticians*, Vol. I, Tables 31 and 32. Cambridge University Press (1954).
29. G. E. P. BOX. *Biometrika*, 40:319 (1953).
30. H. O. HARTLEY. *Communications on Pure and Appl. Math.*, 8:47 (1955).
31. M. KEULS. *Euphytica*, 1:112 (1952).
32. D. B. DUNCAN. *Technometrics*, 7:171 (1965).
33. L. N. BALAAM. *Australian J. Statist.*, 5:62 (1963).
34. H. H. NEWMAN, F. N. FREEMAN, and K. J. HOLZINGER. *Twins*. University of Chicago Press (1937).

Two-way classifications

11.1—Introduction. The experimenter often acquires the ability to predict roughly the behavior of his experimental material. He knows that in identical environments young male rats gain weight faster than young female rats. In a machine which subjects five different pieces of cloth to simulated wearing, he learns from experience that the cloths placed in positions 4 and 5 will receive less abrasion than those in the other positions. Such knowledge can be used to increase the accuracy of an experiment. If there are a treatments to be compared, he first arranges the experimental units in groups of a , often called *replications*. The rule is that units assigned to the same replication should be as similar in responsiveness as possible. Each treatment is then allocated by randomization to one unit in each replication. This produces a two-way classification, since any observation is classified by the treatment which it received and the replication to which it belonged.

Two-way classifications are frequent in surveys also. We already encountered an example (section 9.13) in which farms were classified by soil type and owner-tenant status. In a survey of family expenditures on food, classification of the results by size of family and income level is obviously relevant.

We first present an example to familiarize you with the standard computations needed to perform the analysis of variance and make any desired comparisons. Later, the mathematical assumptions will be discussed.

11.2—An experiment with two criteria of classification. In agricultural experiments the agronomist tries to classify the plots into replications in such a way that soil fertility and growing conditions are as uniform as possible within any replication. In this process he utilizes any knowledge that he has about fertility gradients, drainage, liability to attack by pests, etc. One guiding principle is that, in general, plots that are close together tend to give similar yields. Replications are therefore usually compact areas of land. Within each replication one plot is assigned to each treatment at random. This experimental plan is called *randomized blocks*, the

300 Chapter 11: Two-Way Classifications

replication being a block of land. The two criteria of classification are treatments and replications.

Table 11.2.1 comes from an experiment (1) in which four seed treatments were compared with no treatment (Check) on soybean seeds. The data are the number of plants which failed to emerge out of 100 planted in each plot.

TABLE 11.2.1
ANALYSIS OF VARIANCE OF A 2-WAY CLASSIFICATION
(Number of failures out of 100 planted soybean seeds)

Treatment	Replication					Total	Mean
	1	2	3	4	5		
Check	8	10	12	13	11	54	10.8
Arasan	2	6	7	11	5	31	6.2
Spargon	4	10	9	8	10	41	8.2
Semesan, Jr.	3	5	9	10	6	33	6.6
Fermate	9	7	5	5	3	29	5.8
Total	26	38	42	47	35	188	

Correction. $C = (188)^2/25 = 1,413.76$

Total S.S.: $8^2 + 2^2 + \dots + 6^2 + 3^2 - C = 220.24$

Treatments S.S.: $\frac{54^2 + 31^2 + \dots + 29^2}{5} - C = 83.84$

Replications S.S.: $\frac{26^2 + 38^2 + \dots + 35^2}{5} - C = 49.84$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Replications	4	49.84	12.46
Treatments	4	83.84	20.96
Residuals (Error)	16	86.56	5.41
Total	24	220.24	

The first steps are to find the treatment totals, the replication totals, the grand total, and the usual correction for the mean. The total sum of squares and the sum of squares for Treatments are computed just as in a one-way classification. The new feature is that the sum of squares for Replications is also calculated. The rule for finding this sum of squares is the same as for Treatments. The sum of squares of the replication totals is divided by the number of observations in each replication (5) and the correction factor is subtracted. Finally, in the analysis of variance, we compute the line

$$\text{Residuals} = \text{Total} - \text{Replications} - \text{Treatments}$$

As will be shown later, the Residuals mean square, 5.41, with 16 *df*., is an unbiased estimate of the error variance per observation.

The *F* ratio for treatments is $20.96/5.41 = 3.87$, with 4 and 16 *df*., significant at the 5% level. Actually, since this experiment has certain designed comparisons, discussed in the next section, 11.3, the overall *F*-test is not of great importance. Note that the Replications mean square is more than twice the Residuals mean square. This is an indication of real differences between replication means, suggesting that the classification into replications was successful in improving accuracy. A method of estimating the amount of gain in accuracy will be presented in section 11.7.

EXAMPLE 11.2.1—In three species of citrus trees the ratio of leaf area to dry weight was determined for three conditions of shading (2).

Shading	Shamouti Orange	Marsh Grapefruit	Clementine Mandarin
Sun	112	90	123
Half shade	86	73	89
Shade	80	62	81

Compute the analysis of variance. Ans. Mean squares for shading and error, 942.1 and 21.8. $F = 43.2$, with 2 and 4 *df*. The shading was effective in decreasing the relative leaf area. See example 11.5.4 for further discussion.

EXAMPLE 11.2.2—When there are only two treatments, the data reduce to two paired samples, previously analyzed by the *t*-test in chapter 4. This *t*-test is equivalent to the *F*-test of treatments as given in this section. Verify this result by performing the analysis of variance of the mosaic virus example in section 4.3, p 95, as follows.

	Degrees of Freedom	Sum of Squares	Mean Square
Replications (Pairs)	7	575	82.2
Treatments	1	64	64.0
Error	7	65	9.29

$F = 6.89$, *df* = 1, 7. $\sqrt{F} = 2.63 = t$ as given on p 94

11.3—Comparisons among means. The discussion of different types of comparisons in sections 10.7 and 10.8 applies also to two-way classifications. To illustrate a planned comparison, we compare the mean number of failures for the Check with the corresponding average for the four Chemicals. From table 11.2.1 the means are:

Check 10.8	Arasan 6.2	Spargon 8.2	Semesan, Jr. 6.6	Fermate 5.8
---------------	---------------	----------------	---------------------	----------------

The comparison is, therefore,

$$10.8 - \frac{6.2 + 8.2 + 6.6 + 5.8}{4} = 10.8 - 6.7 = 4.1$$

The experiment has five replications, with $s = \sqrt{5.41} = 2.326$ (16 *d.f.*). Hence, by Rule 10.7.1, the estimated error of the above difference is

$$\frac{s}{\sqrt{5}} \sqrt{1^2 + \frac{1}{4^2} + \frac{1}{4^2} + \frac{1}{4^2} + \frac{1}{4^2}} = \frac{(2.326)}{\sqrt{5}} \sqrt{\frac{5}{4}} \\ = 2.326/2 = 1.163$$

with 16 *d.f.* Thus 95% confidence limits for the average reduction in failure rate due to the Chemicals are

$$4.1 \pm (2.120)(1.163) = 4.1 \pm 2.5, \text{ i.e., } 1.6 \text{ and } 6.6$$

The next step is to compare the means for the four Chemicals. For this, the discussion in section 10.8 is relevant. The *LSD* is

$$t_{0.05} s \sqrt{2/n} = (2.120)(2.326) \sqrt{2/5} = 3.12.$$

Since the largest difference between any two means is $8.2 - 5.8 = 2.4$, there are no significant differences among the Chemicals. You may verify that the Studentized Range *Q*-test requires a difference of 4.21 for significance at the 5% level, giving, of course, the same verdict as the *LSD* test.

11.4—Algebraic notation. For the results of a two-way classification table 11.4.1 gives an algebraic notation that has become standard in mathematical statistics. X_{ij} represents the measurement obtained for the unit that is in the *i*th row (treatment) and *j*th column (replication). Row totals and means are denoted by $X_{i.}$ and $\bar{X}_{i.}$, respectively, while $X_{.j}$ and $\bar{X}_{.j}$ denote column totals and means. The overall mean is $\bar{X}_{..}$. General instructions for computing the analysis of variance appear under the

TABLE 11.4.1
ALGEBRAIC REPRESENTATION OF A 2-WAY TABLE WITH *a* TREATMENTS AND *b* REPLICATIONS
(Computing instructions and analysis of variance)

Treatments $i = 1 \dots a$	Replications, $j = 1 \dots b$					Sum	Mean
	1	...	j	...	b		
1	X_{11}	...	X_{1j}	...	X_{1b}	$X_{1.}$	$\bar{X}_{1.}$
2	X_{21}	...	X_{2j}	...	X_{2b}	$X_{2.}$	$\bar{X}_{2.}$
.
.
i	X_{i1}	...	X_{ij}	...	X_{ib}	$X_{i.}$	$\bar{X}_{i.}$
.
.
a	X_{a1}	...	X_{aj}	...	X_{ab}	$X_{a.}$	$\bar{X}_{a.}$
Sum	$X_{.1}$...	$X_{.j}$...	$X_{.b}$	$X_{..}$	$\bar{X}_{..}$
Mean	$\bar{X}_{.1}$...	$\bar{X}_{.j}$...	$\bar{X}_{.b}$		

TABLE 11.4.1 (Continued)

Correction: $C = (\Sigma X_{ij})^2 / ab = X_{..}^2 / ab$ Total: $\Sigma X_{ij}^2 - C$ Treatments: $A = \frac{X_{1.}^2 + \dots + X_{a.}^2}{b} - C$ Replications: $B = \frac{X_{.1}^2 + \dots + X_{.b}^2}{a} - C$ Residuals: $D = \text{Total} - (\text{Treatments} + \text{Replications})$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Treatments	$a - 1$	A	$A/(a - 1)$
Replications	$b - 1$	B	$B/(b - 1)$
Residuals	$(a - 1)(b - 1)$	D	$D/(a - 1)(b - 1)$
Total	$ab - 1$	$A + B + D$	

table. Note that the number of *d.f.* for Residuals (Error) is $(a - 1)(b - 1)$, the product of the numbers of *d.f.* for rows and columns.

In this book we have kept algebraic symbolism to a minimum, in order to concentrate attention on the data. The symbols are useful, however, in studying the structure of the two-way classification in the next section.

11.5—Mathematical model for a two-way classification. The model being used is

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = 1 \dots a, \quad j = 1 \dots b,$$

where μ represents the overall mean, the α_i stand for fixed row (treatment) effects and the β_j for fixed column (replication) effects. The convention

$$\Sigma \alpha_i = \Sigma \beta_j = 0$$

is usually adopted.

This model involves two basic assumptions:

1. The mathematical form $(\mu + \alpha_i + \beta_j)$ implies that row and column effects are additive. Apart from experimental errors, the difference in effect between treatment 2 and treatment 1 in replication j is

$$(\mu + \alpha_2 + \beta_j) - (\mu + \alpha_1 + \beta_j) = \alpha_2 - \alpha_1$$

This difference is the same in all replications. When we analyze real data, there is no assurance that row and column effects are exactly additive. The additive model is used because of its simplicity and because it is often a good approximation to more complex types of relationships.

2. The ε_{ij} are independent random variables, normally distributed with mean 0 and variance σ^2 . They represent the extent to which the data depart from the additive model because of experimental errors.

304 Chapter 11: Two-Way Classifications

As an aid to understanding the model we shall construct a set of data by its use. Let

$$\mu = 30$$

$$\alpha_1 = 10, \alpha_2 = 3, \alpha_3 = 0, \alpha_4 = -13; \Sigma \alpha_i = 0$$

$$\beta_1 = 1, \beta_2 = -4, \beta_3 = 3; \Sigma \beta_j = 0$$

The ε_{ij} are drawn at random from table 3.2.1, each decreased by 30. This makes the ε_{ij} approximately normal with mean 0 and variance 25.

In each cell of table 11.5.1, $\mu = 30$ is entered first. Next is the treatment α_i , differing from row to row. Following this is the replication effect, one in each column. In each cell, the sum of these three parts is

TABLE 11.5.1
EXPERIMENT CONSTRUCTED ACCORDING TO MODEL I. $\mu = 30$

Treatment	$\beta_1 = 1$	Replication $\beta_2 = -4$	$\beta_3 = 3$	$X_{i.}$	$\bar{X}_{i.}$
$\alpha_1 = 10$	30 10 1 -11 <hr/> $X_{11} = 30$	30 10 -4 -7 <hr/> $X_{12} = 29$	30 10 3 3 <hr/> $X_{13} = 46$	105	35
$\alpha_2 = 3$	30 3 1 1 <hr/> $X_{21} = 35$	30 3 -4 5 <hr/> $X_{22} = 34$	30 3 3 -3 <hr/> $X_{23} = 33$	102	34
$\alpha_3 = 0$	30 0 1 0 <hr/> $X_{31} = 31$	30 0 -4 4 <hr/> $X_{32} = 30$	30 0 3 -1 <hr/> $X_{33} = 32$	93	31
$\alpha_4 = -13$	30 -13 1 -2 <hr/> $X_{41} = 16$	30 -13 -4 -2 <hr/> $X_{42} = 11$	30 -13 3 1 <hr/> $X_{43} = 21$	48	16
$X_{.j}$	112	104	132	348	
$\bar{X}_{.j}$	28	26	33		29
Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square		
Replications	2	104	52		
Treatments	3	702	234		
Residuals	6	132	22		

fixed by μ , the α_i , and the β_j . Sampling variation is introduced by the fourth entry, a deviation drawn at random from table 3.2.1. According to the model, X_{ij} is the sum of the four entries just described.

Some features of the model are now apparent:

(i) The effects of the treatments are not influenced by the β_j because the sum of the β_j in each row is zero. If there were no errors, check from table 11.5.1 that the sum for treatment 1 would be $41 + 36 + 43 = 120$, the mean being $40 = \mu + \alpha_1$. The observed mean, $\bar{X}_1 = 35$, differs from 40 by the mean of the ε_{ij} , namely $(-11 - 7 + 3)/3 = -5$. This is an instance of the general result

$$\bar{X}_i = \mu + \alpha_i + (\varepsilon_{i1} + \varepsilon_{i2} + \dots + \varepsilon_{ib})/b$$

This result shows that \bar{X}_i is an unbiased estimate of $\mu + \alpha_i$ and that its variance is σ^2/b , because the error of the estimate is the mean of b independent errors, each with variance σ^2 .

(ii) In the same way, the replication means are unbiased estimates of $\mu + \beta_j$, with variance σ^2/a .

(iii) In the analysis of variance the Residuals mean square, 22, is an unbiased estimate of $\sigma^2 = 25$. More explanation on this point will be given presently.

(iv) The mean square for Replications is inflated by the β_j and that for Treatments by the α_i . The expected values of these mean squares are shown in table 11.5.2, which deserves careful study. Note that the expected value of the Treatments mean square is the same as in a one-way classification with b observations in each class (compare with equation 10.4.1, p. 265).

TABLE 11 5 2
COMPONENT ANALYSIS OF THE CONSTRUCTED EXPERIMENT

Source of Variation	Degrees of Freedom	Mean Square	Expected Value (Parameters Estimated)
Replications	2	52	$\sigma^2 + a\kappa_B^2$
Treatments	3	234	$\sigma^2 + b\kappa_A^2$
Residuals	6	22	σ^2

$$\kappa_B^2 = \frac{\sum \beta_j^2}{h-1} = \frac{(1)^2 + (-4)^2 + (3)^2}{2} = 13$$

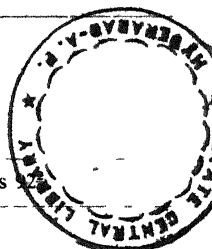
$$\kappa_A^2 = \frac{\sum \alpha_i^2}{a-1} = \frac{(10)^2 + (3)^2 + (0)^2 + (-13)^2}{3} = 92\frac{2}{3}$$

$$s_B^2 = (52 - 22)/4 = 8 \text{ estimates } 13 \quad s_A^2 = (234 - 22)/3 = 71 \text{ estimates } 92\frac{2}{3}$$

Error Mean Square = 22 estimates 25

Replications Mean Square = 52 estimates $25 + 4(13) = 77$

Treatments Mean Square = 234 estimates $25 + 3(93) = 304$



We turn to the estimates of μ , α_i and β_j . These estimates are

$$\hat{\mu} = \bar{X}_{..}; \hat{\alpha}_i = \bar{X}_{i.} - \bar{X}_{..}; \hat{\beta}_j = \bar{X}_{.j} - \bar{X}_{..}$$

If we estimate any individual observation X_{ij} from the fitted model, the estimate is

$$\begin{aligned}\hat{X}_{ij} &= \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{X}_{..} + (\bar{X}_{i.} - \bar{X}_{..}) + (\bar{X}_{.j} - \bar{X}_{..}) \\ &= \bar{X}_{i.} + \bar{X}_{.j} - \bar{X}_{..}\end{aligned}$$

Table 11.5.3 shows the original observations X_{ij} , the estimates \hat{X}_{ij} , and the deviations of the observations from the estimates, $D_{ij} = X_{ij} - \hat{X}_{ij}$. For treatment 1 in replication 2, for instance, we have from table 11.5.1,

$$X_{12} = 29, \hat{X}_{12} = 35 + 26 - 29 = 32, D_{12} = -3$$

TABLE 11 5 3
LINEAR MODEL FITTED TO THE OBSERVATIONS IN TABLE 11 5 1

Treatment		Replication		
		1	2	3
1	X_{ij}	30	29	46
	\hat{X}_{ij}	34	32	39
	D_{ij}	- 4	- 3	+ 7
2	X_{ij}	35	34	33
	\hat{X}_{ij}	33	31	38
	D_{ij}	+ 2	+ 3	- 5
3	X_{ij}	31	30	32
	\hat{X}_{ij}	30	28	35
	D_{ij}	+ 1	+ 2	- 3
4	X_{ij}	16	11	21
	\hat{X}_{ij}	15	13	20
	D_{ij}	+ 1	- 2	+ 1

The deviations D_{ij} have three important properties

- (i) Their sum is zero in any row or column.
- (ii) Their sum of squares,

$$(-4)^2 + (+2)^2 + \quad + (-3)^2 + (+1)^2 = 132,$$

is equal to the Residuals sums of squares in the analysis of variance at the foot of table 11 5 1. Thus the Residuals sum of squares measures the extent to which the linear additive model fails to fit the data. This result is a consequence of a general algebraic identity

$$\begin{aligned}\text{Residuals } S S &= \sum_i \sum_j (X_{ij} - \bar{X}_i - \bar{X}_{.j} + \bar{X}_{..})^2 \\ &= \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 - b \sum_i (\bar{X}_i - \bar{X}_{..})^2 - a \sum_j (\bar{X}_{.j} - \bar{X}_{..})^2 \\ &\quad \text{Total S.S.} \quad - \text{Treatments S.S.} - \text{Replications S.S.}\end{aligned}$$

This equation shows that the analysis of variance is a quick method of finding the sum of squares of the deviations of the observations from the fitted model. When the analysis is programmed for an electronic computer, it is customary to compute and print the D_{ij} . This serves two purposes. It enables the investigator to glance over the D_{ij} for signs of gross errors or systematic departures from the linear model, and it provides a check on the Residuals sum of squares.

(iii) From the constructed model you may verify the remarkable result that

$$D_{ij} = \varepsilon_{ij} - \bar{\varepsilon}_i - \bar{\varepsilon}_{.j} + \bar{\varepsilon}_{..}$$

For example, for treatment 1 in replication 2 you will find from table 11.5.1,

$$\begin{aligned}\varepsilon_{12} &= -7; \bar{\varepsilon}_{1.} = -5; \bar{\varepsilon}_{.2} = 0; \bar{\varepsilon}_{..} = -1 \\ \varepsilon_{12} - \bar{\varepsilon}_{1.} - \bar{\varepsilon}_{.2} + \bar{\varepsilon}_{..} &= (-7) - (-5) - (0) + (-1) = -3,\end{aligned}$$

in agreement with $D_{12} = -3$ in table 11.5.3. Thus, if the additive model holds, each D_{ij} is a linear combination of the random errors. It may be shown that any D_{ij}^2 is an unbiased estimate of $(a-1)(b-1)\sigma^2/ab$. It follows that the Residuals sum of squares is an unbiased estimate of $(a-1)(b-1)\sigma^2$. This gives the basic result that the Residuals mean square, with $(a-1)(b-1) d.f.$, is an unbiased estimate of σ^2 .

To summarize the salient features, the additive model implies that the treatment effects α_i are the same in every replication, and vice versa. If additivity holds (apart from independent errors) the observed treatment means are unbiased estimates of the treatment effects. The F -test may be applied both to Treatments and Replications. The Residuals mean square measures the extent to which the additive model fails to fit the data and provides an unbiased estimate of σ^2 .

EXAMPLE 11.5.1—Suppose that with $a = b = 2$, treatment and replication effects are *multiplicative*. Treatment 2 gives results 20% higher than treatment 1 and replication 2 gives results 10% higher than replication 1. With no random errors, the observations would be as shown on the left below.

Treatment	X_{ij}		Treatment	\hat{X}_{ij}	
	Replication 1	Replication 2		Replication 1	Replication 2
1	1.00	1.10	1	0.995	1.105
2	1.20	1.32	2	1.205	1.315

Verify that the \hat{X}_{ij} given by fitting the linear model are as shown on the right above. Any D_{ij} is only ± 0.005 . The linear model gives a good fit to a multiplicative model when

treatment and replication effects are small or moderate. If, however, treatment 2 gives a 100% increase and replication 2 a 50% increase, you will find $D_{ij} = \pm 0.125$, not so good a fit.

EXAMPLE 11.5.2—In table 11.5.3, verify that $\bar{X}_{33} = 35$ $D_{33} = -3$.

EXAMPLE 11.5.3—Perform an analysis of variance of the \bar{X}_{ij} in table 11.5.3. Verify that the Treatments and Replications sums of squares are the same as for the X_{ij} , but that the Residual sum of squares is zero. Can you explain these results?

EXAMPLE 11.5.4—Calculate the D_{ij} for the 3×3 citrus data in example 11.2.1 and verify that the Residuals mean square, computed from the D_{ij} , is 21.8. Carry one decimal place in the D_{ij} .

EXAMPLE 11.5.5—The result,

$$D_{ij} = \varepsilon_{ij} - \bar{\varepsilon}_{i.} - \bar{\varepsilon}_{.j} + \bar{\varepsilon}_{..},$$

shows that D_{ij} is a linear combination of the form $\Sigma \Sigma \lambda_{ij} \varepsilon_{ij}$. By Rule 10.7.1, its variance is $\sigma^2 \Sigma \Sigma \lambda_{ij}^2$.

For D_{11} , for example, the λ_{ij} work out as follows:

Observations	No. of Terms	λ_{ij}
D_{11}	1	$(a-1)(b-1)/ab$
Rest of D_{1j}	$(b-1)$	$-(a-1)/ab$
Rest of D_{i1}	$(a-1)$	$-(b-1)/ab$
Rest of D_{ij}	$(a-1)(b-1)$	$+1/ab$

It follows that $\Sigma \Sigma \lambda_{ij}^2 = (a-1)(b-1)/ab$. Thus D_{11}^2 and similarly any D_{ij}^2 estimates $(a-1)(b-1)\sigma^2/ab$, as stated in the text.

11.6—Partitioning the treatments sum of squares. When the treatments contain certain planned comparisons, it is often possible to partition the Treatments sum of squares in the analysis of variance in a way that is helpful. Some rules for doing this will now be given. In the analysis of variance, comparisons are usually calculated from the treatment totals T_i rather than the means, since this saves time and avoids rounding errors.

Rule 11.6.1—If $L = \lambda_1 T_1 + \dots + \lambda_a T_a$, ($\Sigma \lambda_i = 0$) is a comparison among the treatment totals, then

$$L^2/n\Sigma\lambda^2$$

is a part of the sum of squares for treatments, associated with a single degree of freedom, where n is the number of observations in any treatment total.

In the experiment on seed treatment of soybeans (table 11.2.1) the comparison Check vs. Chemicals may be represented as follows:

	Check	Arasan	Spargon	Semesan, Jr.	Fermate
Total (T_i)	54	31	41	33	29
λ_i	4	-1	-1	-1	-1

To avoid fractions the λ_i have been taken as 4, -1, -1, -1, -1 instead of as 1, -1/4, -1/4, -1/4, -1/4 in section 11.3. This gives

$$L = 4(54) - 31 - 41 - 33 - 29 = 82$$

Since $n = 5$, the contribution to the Treatments sum of squares is

$$L^2/n\Sigma\lambda^2 = (82)^2/(5)(20) = 67.24 \quad (1 \text{ d.f.})$$

The Treatments sum of squares was 83.84 with 4 d.f. The remaining part is therefore 16.60 with 3 d.f. What does it represent? As might be guessed, it represents the sum of squares of Deviations of the totals for the four Chemicals from their mean, namely,

$$\frac{31^2 + 41^2 + 33^2 + 29^2}{5} - \frac{134^2}{20} = 16.60$$

Thus, the original analysis of variance in table 11.2.1 might be reported as follows:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Check vs. Chemicals	1	67.24	67.24
Among Chemicals	3	16.60	5.53
Residuals (Error)	16	86.56	5.41

The F ratio $67.24/5.41 = 12.43$ ($P < 0.01$) shows that the average failure rates are different for Check and Chemicals (though as usual it does not tell us the size and direction of the effect in terms of means). The F ratio $5.53/5.41 = 1.02$ for Among Chemicals warns us that there are unlikely to be any significant differences among Chemicals, as was already verified.

As a second example, consider the data on the effect of shade on the ratio of leaf area to leaf weight in citrus trees (example 11.2.1). The "treatment" totals, $n = 3$, were as follows:

Totals	T_i	Sun 325	Half Shade 248	Shade 223	Comparison L_i	Divisor	S.S. = $L^2/n\Sigma\lambda^2$
Effect of shade	λ_{1i}	+1	0	-1	102	6	1734
Half shade vs. Rest	λ_{2i}	+1	-2	+1	52	18	150

We might measure the effect of shade by the extreme comparison $L_1 = (\text{Sun} - \text{Shade})$. We might also be interested in whether the results for Half Shade are the simple average of those for Sun and Shade. This gives the comparison L_2 .

Rule 11.6.2—Two comparisons:

$$\begin{aligned} L_1 &= \lambda_{11}T_1 + \lambda_{12}T_2 + \dots + \lambda_{1a}T_a = \Sigma\lambda_{1i}T_i, \\ L_2 &= \lambda_{21}T_1 + \lambda_{22}T_2 + \dots + \lambda_{2a}T_a = \Sigma\lambda_{2i}T_i, \end{aligned}$$

are orthogonal if

$$\lambda_{11}\lambda_{21} + \lambda_{12}\lambda_{22} + \dots + \lambda_{1a}\lambda_{2a} = 0 \quad : \text{ i.e. } \Sigma\lambda_{1i}\lambda_{2i} = 0$$

In applying this rule, if a total T_i does not enter into a comparison, its coefficient is taken as zero.

The comparisons L_1 and L_2 are orthogonal, since

$$(+1)(+1) + (0)(-2) + (-1)(+1) = 0$$

Rule 11.6.3—If two comparisons are orthogonal, their contributions $L_1^2/n\Sigma\lambda_{1i}^2$ and $L_2^2/n\Sigma\lambda_{2i}^2$ are independent parts of the sum of squares for treatments, each with 1 *d.f.*

This means that the Treatments S.S. may be partitioned into the contributions due to L_1 and L_2 , plus any remainder (with $(a - 3)$ *d.f.*). A consequence of this rule is

Rule 11.6.4—Among a treatments, if $(a - 1)$ comparisons are mutually orthogonal (i.e., every pair is orthogonal), then

$$\frac{L_1^2}{n\Sigma\lambda_{1i}^2} + \frac{L_2^2}{n\Sigma\lambda_{2i}^2} + \dots + \frac{L_{a-1}^2}{n\Sigma\lambda_{(a-1)i}^2} = \text{Treatments S.S.}$$

The citrus data, with $a = 3$, are an example. The sum of the squared contributions for L_1 and L_2 is $1734 + 150 = 1884$, which may be verified to be the Treatments S.S. Thus, the relevant part of the analysis of variance can be presented as follows:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	<i>F</i>
Effect of shade	1	1734	1734	79.5
Half shade vs. Rest	1	150	150	6.9
Error	4	87	21.8	

The F value for the effect of shade is highly significant. With 1 and 4 *d.f.*, $F = 6.9$ for the comparison of half shade with the average of sun and shade does not quite reach the 5% level. There is a suggestion, however, that the results for half shade are closer to those for shade than to those for sun. Both these comparisons can, of course, be examined by t -tests on the treatment means.

EXAMPLE 11 6.1—In the following artificial example, two of the treatments were variants of one type of process, while the other four were variants of a second type. The treatment totals (4 replications) were.

Process 1			Process 2		
59	68	70	84	76	81

Partition the Treatments S.S. as follows

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Between processes	1	67.69	67.69
Variants of process 1	1	10.12	10.12
Variants of process 2	3	28.19	9.40

11.7—Efficiency of blocking. When an experiment has been set out in replications, using the randomized blocks design, it is sometimes of interest to know how effective the blocking was in increasing the precision of the comparisons, particularly if there is doubt whether the criterion used in constructing the replications is a good one, or if the use of these replications is troublesome. From the analysis of variance of a randomized blocks experiment, we can estimate the error variance that would have been obtained if a completely random arrangement of the same experimental units (plots) had been used instead of randomized blocks.

Call the two error variances s_{CR}^2 and s_{RB}^2 . With randomized blocks the variance of a treatment mean is s_{RB}^2/b . To get the same variance of a treatment mean with complete randomization, the number of replications n must satisfy the relation

$$\frac{s_{CR}^2}{n} = \frac{s_{RB}^2}{b} \quad \text{or} \quad \frac{n}{b} = \frac{s_{CR}^2}{s_{RB}^2}$$

For this reason the ratio s_{CR}^2/s_{RB}^2 is used to measure the *relative efficiency* of the blocking.

If M_B and M_E are the mean squares for blocks and error in the analysis of variance of randomized blocks experiment that has been performed, it has been shown (3, 4) that

$$\frac{s_{CR}^2}{s_{RB}^2} = \frac{(b-1)M_B + b(a-1)M_E}{(ab-1)M_E}$$

Using the soybeans experiment as an example (table 11.2.1), $M_B = 12.46$, $M_E = 5.41$, $a = b = 5$,

$$\frac{s_{CR}^2}{s_{RB}^2} = \frac{4(12.46) + 20(5.41)}{24(5.41)} = 1.22$$

With complete randomization, about six replications instead of five would have been necessary to obtain the same standard error of a treatment mean.

This comparison is not quite fair to complete randomization, which would provide 20 *d.f.* for error as against 16 with randomized blocks and therefore require smaller values of t in calculating confidence intervals. This is taken into account by a formula suggested by Fisher (5), which replaces the ratio s_{CR}^2/s_{RB}^2 by the following ratio:

$$\begin{aligned} \text{Relative amount of information} &= \frac{(f_{RB} + 1)(f_{CK} + 3)}{(f_{RB} + 3)(f_{CR} + 1)} \frac{s_{CR}^2}{s_{RB}^2} \\ &= \frac{(16 + 1)(20 + 3)}{(16 + 3)(20 + 1)} (1.22) = 1.20 \end{aligned}$$

The adjustment for *d f* has little effect here but makes more difference in small experiments

EXAMPLE 11.7.1—In a randomized-blocks experiment which compared four strains of Gallipoli wheat (6) the mean yields (pounds per plot) and the analysis of variance were as follows:

Strain	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
Mean yield	34.4	34.8	33.7	28.4
Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	
Blocks	4	21.46		5.36
Strains	3	134.45		44.82
Error	12	26.26		2.19

(i) How many replications were there? (ii) Estimate s_{CR}^2/s_{RB}^2 . (iii) Estimate the relative amount of information by Fisher's formula. Ans. (i) 1.30, (iii) 1.26

EXAMPLE 11.7.2—In example 11.7.1, verify that the *LSD* and the *Q* methods both show *D* inferior to the other strains, but reveal no differences among the other strains.

11.8—Latin squares. In agricultural field experiments, there is frequently a gradient in fertility running parallel to one of the sides of the field. Sometimes, gradients run parallel to both sides and sometimes, in a new field, it is not known in which direction the predominant gradient may run. A useful plan for such situations is the Latin square. With four treatments, *A*, *B*, *C*, *D*, it may be like this:

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>C</i>	<i>A</i>	<i>D</i>	<i>B</i>
<i>D</i>	<i>C</i>	<i>B</i>	<i>A</i>
<i>B</i>	<i>D</i>	<i>A</i>	<i>C</i>

The rows and columns of the square are parallel to the two sides of the field. Each treatment appears once in every row and once in every column, this being the basic property of a Latin square. Differences in fertility between rows and differences between columns are both eliminated from the comparison of the treatment means, with a resultant increase in the precision of the experiment.

In numerous other situations the Latin square is also effective in controlling two sources of variation of which the investigator has predictive knowledge. In psychology and medicine, the human subject frequently comprises a replication of the experiment, receiving all the treatments in succession, with intervening intervals in which the effects of previous treatment will have died away. However, a systematic effect of the order in which the treatments are given can often be detected. This is controlled by making the columns of the square represent the order, while rows represent subjects. In animal nutrition, the effects of both litter and condition of the animal may be removed from the estimates of treatment means by the use of a Latin square.

To construct a Latin square, write down a systematic arrangement

of the letters and rearrange rows and columns at random. Then assign treatments at random to the letters. For refinements, see (7).

The model for a Latin square experiment (model I) is

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}; \quad i, j \text{ and } k = 1 \dots a; \quad \varepsilon_{ijk} = \mathcal{N}(0, \sigma)$$

where α , β , and γ indicate treatment, row, and column effects, with the usual convention that their sums are zero. The assumption of additivity is carried a step further than with a two-way classification, since we assume the effects of all three factors to be additive.

It follows from the model that a treatment mean $\bar{X}_{i..}$ is an unbiased estimate of $\mu + \alpha_i$, the effects of rows and columns canceling out because of the symmetry of the design. The standard error of $\bar{X}_{i..}$ is σ/\sqrt{a} . The estimate \hat{X}_{ijk} of the observation X_{ijk} made from the fitted linear model is

$$\hat{X}_{ijk} = \bar{X}... + (\bar{X}_{i..} - \bar{X}...) + (\bar{X}_{.j.} - \bar{X}...) + (\bar{X}_{..k} - \bar{X}...)$$

Hence, the deviation from the fitted model is

$$D_{ijk} = X_{ijk} - \hat{X}_{ijk} = X_{ijk} - \bar{X}_{i..} - \bar{X}_{.j.} - \bar{X}_{..k} + 2\bar{X}...$$

As in the two-way classification, the error sum of squares in the analysis of variance is the sum of the D_{ijk}^2 and the Error mean square is an unbiased estimate of σ^2 .

Table 11.8.1 shows the field layout and yields of a 5×5 Latin square experiment on the effects of spacing on yields of millet plants (8). In the computations, the sums for rows and columns are supplemented by sums

TABLE 11.8.1
YIELDS (GRAMS) OF PLOTS OF MILLET ARRANGED IN A LATIN SQUARE
(Spacings: A, 2-inch; B, 4; C, 6; D, 8; E, 10)

Row	Column					Sum
	1	2	3	4	5	
1	B: 257	E: 230	A: 279	C: 287	D: 202	1,255
2	D: 245	A: 283	E: 245	B: 280	C: 260	1,313
3	E: 182	B: 252	C: 280	D: 246	A: 250	1,210
4	A: 203	C: 204	D: 227	E: 193	B: 259	1,086
5	C: 231	D: 271	B: 266	A: 334	E: 338	1,440
Sum	1,118	1,240	1,297	1,340	1,309	6,304
Summary by Spacing						
	A 2"	B 4"	C 6"	D 8"	E 10"	
Sum	1,349	1,314	1,262	1,191	1,188	6,304
Mean	269.8	262.8	252.4	238.2	237.6	252.2

(Continued next page)

TABLE 11 8 1 (Continued)

Correction	$(6,304)^2/25 = 1,589,617$			
Total	$(257)^2 + \quad + (338)^2 - 1,589,617 = 36,571$			
Rows	$(1,255)^2 + \quad + (1,440)^2$	5	$- 1,589,617 = 13,601$	
Columns	$(1,118)^2 + \quad + (1,309)^2$	5	$- 1,589,617 = 6,146$	
Spacings	$(1,349)^2 + \quad + (1,188)^2$	5	$- 1,589,617 = 4,156$	
Error	12,668			

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Total	24	36,571	
Rows	4	13,601	3,400
Columns	4	6,146	1,536
Spacings	4	4,156	1,039
Error	12	12,668	1,056

and means for treatments (spacings). By the usual rules, sums of squares for Rows, Column, and Spacings are calculated. These are subtracted from the Total SS to give the Error SS , with $(a-1)(a-2) = 12 df$.

Table 11 8 2 shows the expected values of the mean squares, with the usual notation. For illustration we have presented the results that apply if the β_j and γ_k in rows and columns represent random effects, with fixed treatment effects α_i .

TABLE 11 8 2
COMPONENT ANALYSIS IN LATIN SQUARE

Source of Variation	Degrees of Freedom	Mean Square	Estimates of
Rows, R	$a - 1$	M_R	$\sigma^2 + a\sigma_R^2$
Columns, C	$a - 1$	M_C	$\sigma^2 + a\sigma_C^2$
Treatments, A	$a - 1$	M_A	$\sigma^2 + a\kappa_A^2$
Error	$(a-1)(a-2)$	M_E	σ^2

This experiment is typical of many in which the treatments consist of a series of levels of a variable, in this case width of spacing. The objective is to determine the relation between the treatment mean yields, which we will now denote by \bar{Y}_i , and width of spacing X_i . Inspection of the mean yields suggests that the relation may be linear, the yield decreasing steadily as spacing increases. The X_i , x_i , and \bar{Y}_i , are shown in table 11 8 3.

TABLE 11 8 3
DATA FOR CALCULATING THE REGRESSION OF YIELD ON SPACING

Spacing, X_i	2	4	6	8	10
$x_i = X_i - \bar{X}$	-4	-2	0	2	4
\bar{Y}_i (gms)	269.8	262.8	252.4	238.2	237.6

The regression coefficient of yield on spacing is

$$b = \frac{\Sigma(X_i - \bar{X})(\bar{Y}_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} = \frac{\Sigma x_i \bar{Y}_i}{\Sigma x_i^2} = -\frac{178.0}{40} = -4.45,$$

the units being grams per inch increase in spacing. Notice that b is a comparison among the treatment means, with $\lambda_i = x_i$, Σx_i^2 . From Rule 10.7.1, the standard error of b is

$$s_b = \sqrt{(s^2 \Sigma \lambda^2 / a)} = \sqrt{(s^2 / a \Sigma x^2)} = \sqrt{\{(1056) / (5)(40)\}} = 2.298$$

With 12 *df*, 95% confidence limits for the population regression are +0.6 and -9.5 grams per inch increase. The linear decrease in yield is not quite significant, since the limits include 0.

In the analysis of variance, the Treatments *SS* can be partitioned into a part representing the linear regression on width of spacing and a part representing the deviations of the treatment means from the linear regression. This partition provides new information. If the true regression of the means on width of spacing is linear, the Deviations mean square should be an estimate of σ^2 . If the true regression is curved, the Deviations mean square is inflated by the failure of the fitted straight line to represent the curved relationship. Consequently, $F = \text{Deviations} / M S / \text{Error}$. *M S* tests whether the straight line is an adequate fit.

The sum of squares for Regression (1 *df*) can be computed by the methods on regression given in chapter 6. In section 6.15 (p. 162) this sum of squares was presented as $(\Sigma xy)^2 / \Sigma x^2$ (table 6.15.3). In this example we have already found $\Sigma xy = \Sigma x_i \bar{Y}_i = -178.0$, and $\Sigma x^2 = 40$, giving $(\Sigma xy)^2 / \Sigma x^2 = (178.0)^2 / 40 = 792.1$. Since, however, each \bar{Y}_i is the mean of five observations, we multiply by 5 when entering this term in the analysis of variance, giving 3,960. The *SS* for Deviations from the regression is found by subtracting 3,960 from the total *SS* for Spacings, 4156 (table 11.8.4).

TABLE 11.8.4
ANALYSIS OF REGRESSION OF SPACING MEAN ON WIDTH OF SPACING
(Millet experiment)

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	<i>F</i>
Spacings (table 11.8.1)	4	4156		
{ Regression	{ 1	3960	3960	3.75
{ Deviations	{ 3	196	66	0.06
Error (table 11.8.1)	12	12668	1056	

The *F*-ratio for Deviations is very small, 0.06, giving no indication that the regression is curved. The *F* for Regression, 3.75, is not quite significant, this test being the same as the *t*-test for b .

The results of this experiment are probably disappointing. In trying to discover the best width of spacing, an investigator hopes to obtain a

curved regression, with reduced yields at the narrowest and widest spacings, so that his range of spacings straddles the optimum. As it is, assuming the linear regression real, the best spacing may lie below 2 in. Methods of dealing with curved regressions in the analysis of variance are given in chapter 12.

Since the number of replications in the Latin square is equal to the number of treatments, the experimenter is ordinarily limited to eight or ten treatments if he uses this design. For four or less treatments, the degrees of freedom for error are fewer than desirable, $(a-1)(a-2) = (3)(2) = 6$ for the 4×4 . This difficulty can be remedied by replicating the squares.

The relative efficiency of a Latin square experiment as compared to complete randomization is

$$\frac{M_R + M_C + (a-1)M_E}{(a+1)M_E}$$

Substituting the millet data:

$$\text{Relative Efficiency} = \frac{s_{CR}^2}{s_L^2} = \frac{3400 + 1536 + (5-1)(1056)}{(5+1)(1056)} = 145\%,$$

a gain of 45% over complete randomization.

There may be some interest in knowing the relative efficiency as compared to a randomized blocks experiment in which either rows or columns were lacking. In the millet experiment since the column mean square was small (this may have been an accident of sampling), it might have been omitted and the rows retained as blocks. The relative efficiency of the Latin square is

$$\frac{M_C + (a-1)M_E}{aM_E} = \frac{1536 + (5-1)1056}{(5)(1056)} = 109\%$$

Kemphorne (4) reminds us that this may not be a realistic comparison. For the blocks experiment the shape of the plots would presumably have been changed, improving the efficiency of that experiment. In this millet experiment, appropriately shaped plots in randomized blocks might well have compensated for the column control.

EXAMPLE 11.8.1—Here is a Latin square for easy computation. Treatments are indicated by A, B, and C.

Rows	Columns		
	1	2	3
1	B: 23	A: 17	C: 29
2	A: 16	C: 25	B: 16
3	C: 24	B: 18	A: 12

The mean squares are: rows, 21; columns, 3; treatments, 93; remainder, 3.

EXAMPLE 11.8.2—Fit the linear model for Latin squares to the data of example 11.8.1. Verify the fitting by the relation, $\sum D_{ijk}^2 = 6$.

EXAMPLE 11.8.3—In experiments affecting the milk yield of dairy cows the great variation among individuals requires large numbers of animals for evaluating moderate differences. Efforts to apply several treatments successively to the same cow are complicated by the decreasing milk flow, by the shapes of the lactation curves, by carry-over effects, and by presumed correlation among the errors, ε_{ijk} . The effort was made to control these difficulties by the use of several pairs of orthogonal Latin squares (9), the columns representing cows, the rows successive periods during lactation, the treatments being A = roughage, B = limited grain, C = full grain.

For this example, a single square is presented, no effort being made to deal with carry-over effects. The entries are pounds of milk for a 6-week period. Compute the analysis of variance.

Period	Cow		
	1	2	3
I	A: 608	B: 885	C: 940
II	B: 715	C: 1087	A: 766
III	C: 844	A: 711	B: 832

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Periods	2	5,900	2,950
Cows	2	47,214	23,607
Treatments	2	103,436	51,718
Error	2	4,843	2,422

11.9—Missing data. Accidents often result in the loss of data. Crops may be destroyed, animals die, or errors made in the application of the treatments or in recording. Although the least squares procedure can be applied to the data that are present, missing items destroy the symmetry and simplicity of the analysis. The calculational methods that have been presented cannot be used. Fortunately, the missing data can be estimated by least squares and entered in the vacant cells of the table. Application of the usual analysis of variance, with some modifications, then gives results that are correct enough for practical purposes.

In these methods the missing items must not be due to failure of a treatment. If a treatment has killed the plants, producing zero yield, this should be entered as 0, not as a missing value.

In a one-way classification (complete randomization) the effect of missing values is merely to reduce the sample sizes in the affected classes. The analysis is handled correctly by the methods for one-way classifications with unequal numbers (section 10.12). No substitution of the missing data is required.

In randomized blocks, a single missing value is estimated by the formula (26)

$$X = \frac{aT + bB - S}{(a-1)(b-1)},$$

where

a = number of treatments

b = number of blocks

T = sum of items with same treatment as missing item

B = sum of items in same block as missing item

S = sum of all observed items

As an example, table 11.9.1 shows the yields in an experiment on four strains of Gallipoli wheat, in which we have supposed that the yield for strain D in block 1 is missing. We have

$$T = 112.6, B = 96.4, S = 627.1, a = 4, b = 5.$$

$$X = \frac{4(112.6) + (5)(96.4) - 627.1}{(3)(4)} = 25.4 \text{ pounds}$$

TABLE 11.9.1
YIELDS OF FOUR STRAINS OF WHEAT IN FIVE RANDOMIZED BLOCKS
(POUNDS PER PLOT) WITH ONE MISSING VALUE

Strain	1	2	Block 3	4	5	Total
A	32.3	34.0	34.3	35.0	36.5	172.1
B	33.3	33.0	36.3	36.8	34.5	173.9
C	30.8	34.3	35.3	32.3	35.8	168.5
D	26.0	29.8	28.0	28.8	112.6
Total	96.4	127.3	135.7	132.1	135.6	627.1

Analysis of Variance (With 25.4 Inserted)

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares
Blocks	4	35.39	
Strains	3	171.36	57.12 (45.79)
Error	12	17.33	1.58
Total	19	224.08	

This value is entered in the table as the yield of the missing plot. All sums of squares in the analysis of variance are then computed as usual. However, the degrees of freedom in the Total and Error $S.S.$, are both reduced by 1, since there are actually only 18 $d.f.$ for the Total $S.S.$ and 11 for Error.

This method gives the correct least squares estimates of the treatment means and of the Error mean square. For the comparison of treatment means, the $s.e.$ of the difference between the mean with a missing value and another treatment mean is not $\sqrt{(2s^2/b)}$ but the larger quantity

$$\sqrt{s^2 \left[\frac{2}{b} + \frac{a}{b(b-1)(a-1)} \right]} = \sqrt{(1.58) \left[\frac{2}{5} + \frac{4}{(5)(4)(3)} \right]} = \pm 0.859,$$

as against ± 0.795 for a pair of treatments with no missing values.

The Treatments (Strains) mean square in the analysis of variance is slightly inflated. The correction for this upward bias is to subtract from the *mean square*

$$\frac{\{B - (a-1)X\}^2}{a(a-1)^2} = \frac{\{96.4 - (3)(25.4)\}^2}{(4)(3)(3)} = 11.33$$

This gives $57.12 - 11.33 = 45.79$ for the correct mean square.

This analysis does not in any sense recover the lost information, but makes the best of what we have.

For the Latin square the formulas are:

$$X = [a(R + C + T) - 2S]/(a-1)(a-2)$$

Deduction from Treatments *mean square* for bias

$$= [S - R - C - (a-1)T]^2/(a-1)^3(a-2)^2$$

where a is the number of treatments, rows, or columns.

To illustrate, suppose that in example 11.8.3 the milk yield, 608 pounds, for Cow 1 in Period I was missing. Table 11.9.2 gives the resulting data and analysis. The correct Treatments mean square is (40, 408).

$$X = \frac{3(1825 + 1559 + 1477) - 2(6780)}{(2)(1)} = 512 \text{ pounds}$$

$$\text{Bias} = \frac{[6780 - 1825 - 1559 - (2)(1477)]^2}{(2)(2)(2)(1)(1)} = 24,420$$

TABLE 11.9.2
3 × 3 LATIN SQUARE WITH ONE MISSING VALUE

Period	1	Cow 2	3	Total	Treatments
I	A ...	B 885	C 940	1,825	A 1,477
II	B 715	C 1,087	A 766	2,568	B 2,432
III	C 844	A 711	B 832	2,387	C 2,871
Total	1,559	2,683	2,538	6,780	6,780
Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares		
Rows (Periods)	2	9,847			
Columns (Cows)	2	68,185			
Treatments	2	129,655	64,828 (40,408)		
Error	1	2,773	2,773		
Total	7	210,460			

Of course, no worthwhile conclusions are likely to flow from a single 3×3 square with a missing value, the Error M.S. having only 1 *d.f.* The *s.e.* of the difference between the treatment mean with the missing value and any other treatment mean is

$$\sqrt{s^2 \left\{ \frac{2}{a} + \frac{1}{(a-1)(a-2)} \right\}}$$

Two or more missing data require more complicated methods. But for a few missing values an iterative scheme may be used for estimation.

To illustrate the iteration, the data in table 11.9.3 seem adequate. Start by entering a reasonable value for one of the missing data, say $X_{22} = 10.5$. This could be $\bar{X}_{..} = 9.3$, but both the block and treatment means are above average, so 10.5 seems better. From the formula, X_{31} is

$$X_{31} = \frac{(3)(27) + (3)(21) - 75.5}{(3-1)(3-1)} = 17.1$$

Substituting $X_{31} = 17.1$ in the table, try for a better estimate of X_{22} by using the formula for X_{22} missing:

$$X_{22} = \frac{(3)(23) + (3)(20) - 82.1}{4} = 11.7$$

With this revised estimate of X_{22} , re-estimate X_{31} :

$$X_{31} = \frac{(3)(27) + (3)(21) - 76.7}{4} = 16.8$$

Finally, with this new value of X_{31} in the table, calculate $X_{22} = 11.8$. One stops because with $X_{22} = 11.8$ no change occurs when X_{31} is recalculated.

In the analysis of variance, subtract 2 *d.f.* from the Total and Error sums of squares. The Treatments *S.S.* and *M.S.* are biased upwards. To obtain the correct Treatments *S.S.*, reanalyze the data in table 11.9.3, ignoring the treatments and the missing values, as a one-way classification with unequal numbers, the blocks being the classes. The new Error (Within blocks) *S.S.* will be found to be 122.50 with 4 *d.f.* Subtract from this the Error *S.S.* that you obtained in the randomized blocks analysis of the completed data. This is 6.40, with 2 *d.f.* The difference, $122.50 - 6.40 = 116.10$, with $4 - 2 = 2$ *d.f.* is the correct Treatments *S.S.* The *F* ratio is $58.05/3.20 = 18.1$, with 2 and 2 *d.f.*

The same method applies to a Latin square with two missing values, with repeated use of the formula for inserting a missing value in a Latin square. Formulas needed for confidence limits and *t*-tests involving the treatment means are given in (3). For experiments analyzed by electronic computers, a general method of estimating missing values is presented in (10)

TABLE 11.9.3
RANDOMIZED BLOCKS EXPERIMENT WITH TWO MISSING VALUES

Treatments	Blocks			Sums
	1	2	3	
<i>A</i>	6	5	4	15
<i>B</i>	15	X_{22}	8	23
<i>C</i>	X_{31}	15	12	27
Sums	21	20	24	65

11.10—Non-conformity to model. In the standard analyses of variance the model specifies that the effects of the different fixed factors (treatments, row, columns, etc.) are additive, and that the errors are normally and independently distributed with the same variance. It is unlikely that these ideal conditions are ever exactly realized in practice. Much research has been done to investigate the consequences of various types of failure in the assumptions; for an excellent review, see (11). Minor failures do not greatly disturb the conclusions drawn from the standard analysis. In subsequent sections some advice is given on the detection and handling of more serious failures. For this discussion the types of failure are classified into gross errors, lack of independence of errors, unequal error variances due to the nature of the treatments, non-normality of errors, and non-additivity.

11.11—Gross errors: rejection of extreme observations. A measurement may be read, recorded, or transcribed wrongly, or a mistake may be made in the way in which the treatment was applied for this measurement. A major error greatly distorts the mean of the treatment involved, and, by inflating the error variance, affects conclusions about the other treatments as well. The principal safeguards are vigilance in carrying out the operating instructions for the experiment and in the measuring and recording process, and eye inspection of the data.

If a figure in the data to be analyzed looks suspicious, an inquiry about this observation sometimes shows that there was a gross error and may also reveal the correct value for this observation. (One should check that the same source of error has not affected other observations also.) With two-way and Latin square classifications, it is harder to spot an unusual observation in the original data, because the expected value of any observation depends on the row, column, and treatment effects. Instead, look at the residuals of the observations from their expected values. In the two-way classification, the residual D_{ij} is

$$D_{ij} = X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}..$$

while in the Latin square,

$$D_{ijk} = X_{ijk} - \bar{X}_{i..} - \bar{X}_{.j.} - \bar{X}_{..k} + 2\bar{X}...$$

If no explanation of an extreme residual that enables it to be corrected is discovered, we may consider rejecting it and analyzing the data by the method in section 11.9 for results with missing observations. The discussion of rules for the rejection of observations began well over a century ago in astronomy and geodesy. Most rules have been based on something like a test of significance. The investigator computes the probability that a residual as large as the suspect would occur by chance if there is no gross error (taking account of the fact that the *largest* residual was selected). If this probability is sufficiently small, the suspect is rejected.

Anscombe (12) points out that it may be wiser to think of a rejection rule as analogous to an insurance policy on a house or an automobile. We pay a premium to protect us against the possibility of damage. In considering whether a proposed policy is attractive, we take into account the size of the premium, our guesses as to the probability that damage will occur, and our estimate of the amount of likely damage if there is a mishap.

A premium is involved in a rejection rule because any rule occasionally rejects an observation that is not a gross error. When this happens, the mean of the affected treatment is less accurately estimated than if we had not applied the rule. If these erroneous rejections cause the variances of the estimated treatment means to be increased by $P\%$, on the average over repeated applications, the rule is said to have a premium of $P\%$.

Anscombe and Tukey (13) present a rule that rejects an observation whose residual has the value d if $|d| > Cs$, where C is a constant to be determined and s is the *S.D.* of the experimental errors (square root of the Error or Residuals mean square). For any small value of P , say $2\frac{1}{2}\%$ or 5% , an approximate method of computing C is given (13). This method applies to the one-way, two-way, and Latin square classifications, as well as to other standard classifications with replication. The formula for C involves the number of Error *d.f.*, say f , and the total number of residuals, say N . In our notation the values of f and N are as follows:

Classification

One-way (a classes, n per class). $f = a(n - 1): N = an$

Two-way (a rows, b columns). $f = (a - 1)(b - 1): N = ab$

Latin square ($a \times a$) $f = (a - 1)(a - 2): N = a^2$

The formula has three steps:

1. Find the *one-tailed* normal deviate z corresponding to the probability $fP/100N$, where P is the premium expressed in per cents.
2. Calculate $K = 1.40 + 0.85z$

$$3. C = K \left\{ 1 - \frac{K^2 - 2}{4f} \right\} \sqrt{\frac{f}{N}}$$

In order to apply this rule, first analyze the data and obtain the values of d and s . To illustrate, consider the randomized blocks wheat data (table 11.9.1, p. 318) with $a = 4$, $b = 5$, that was used as an example of a

missing observation. This observation, for Strain *D* in Block 1, was actually present and had a value 29.3. In the analysis of the complete data, this observation gave the largest residual, 2.3, of all $N = 20$ observations. For the complete data, $s = 1.48$ with $f = 12$. In a rejection rule with a $2\frac{1}{2}\%$ premium, would this observation be rejected?

Since $N = 20$, we have $f/N = 0.6$, $P = 2.5$, so that $fP/100N = (0.6) \cdot (0.025) = 0.015$. From the normal table, this gives $z = 2.170$. Thus,

$$K = 1.40 + (0.85)(2.170) = 3.24$$

$$C = 3.24 \left\{ 1 - \frac{8.50}{48} \right\} \sqrt{0.6} = 2.07$$

Since $Cs = (2.07)(1.48) = 3.06$, a residual of 2.3 does not call for rejection.

EXAMPLE 11.11.1—In the 5×5 Latin square on p. 313, the largest residual from the fitted model is +55.0 for treatment *E* in row 5 and column 5. Would this observation be rejected in a policy with a 5% premium? Ans. No. $Cs = 58.5$

11.12—Lack of independence in the errors. If care is not taken, an experiment may be conducted in a way that induces positive correlations between the errors for different replicates of the same treatment. In an industrial experiment, all the replications of a given treatment might be processed at the same time by the same technicians, in order to cut down the chance of mistakes or to save money. Any differences that exist between the batches of raw materials used with different treatments or in the working methods of the technicians may create positive correlations within treatments.

In the simplest case these situations are represented mathematically by supposing that there is an intraclass correlation ρ_I between any pair of errors within the same treatment. In the absence of real treatment effects, the mean square between treatments is an unbiased estimate of $\sigma^2\{1 + (n - 1)\rho_I\}$, where n is the number of replications, while the error mean square is an unbiased estimate of $\sigma^2(1 - \rho_I)$, as pointed out in section 10.20. The F -ratio is an estimate of $\{1 + (n - 1)\rho_I\}/(1 - \rho_I)$. With ρ_I positive, this ratio can be much larger than 1; for instance, with $\rho_I = 0.2$ and $n = 6$, the ratio is 2.5. Thus, positive correlations among the errors within a treatment vitiate the F -test, giving too many significant results. The disturbance affects t -tests also, and may be major.

In more complex situations the consequences of correlations among the errors have not been adequately studied, but there is reason to believe that they can be serious. Such correlations often go unnoticed, because their presence is difficult to detect by inspection of the data. The most effective precaution is the skillful use of randomization (section 4.12). If it is suspected that observations made within the same time period (e.g., morning or day) will be positively correlated, the order of processing of the treatments within a replication should be randomized. A systematic pattern of errors, if detected, can sometimes be handled by constructing an

appropriate model for the statistical analysis. For examples, see (14), (15), and (16).

11.13—Unequal error variances due to treatments. Sometimes one or more treatments have variances differing from the rest, although there is no reason to suspect non-normality of errors. If the treatments consist of different amounts of lime applied to acid soil, the smallest dressings might give uniformly low yields with a small variance, while the highest dressings, being large enough to overcome the acidity, give good yields with a moderate variance. Intermediate dressings might give good yields on some plots and poor yields on others, and thus show the highest variance. Another example occurs in experiments in which the treatments represent different measuring instruments, some highly precise and some cruder and less expensive. The average readings given by different instruments are being compared in order to check whether the inexpensive instruments are biased. Here we would obviously expect the variance to differ from instrument to instrument.

When the error variance is heterogeneous in this way, the F -test tends to give too many significant results. This disturbance is usually only moderate if every treatment has the same number of replications (11). Comparison of pairs or sub-groups of treatment means may, however, be seriously affected, since the usual estimate of error variance, which pools the variance over all treatments, will give standard errors that are too large for some comparisons and too small for others.

For any comparison $\Sigma \lambda_i \bar{X}_i$ among the class means in a one-way classification, an unbiased estimate of its error variance is $V = \Sigma \lambda_i^2 s_i^2 / n_i$, where n_i is the number of replications in \bar{X}_i and s_i^2 is the mean square within the i th class. This result holds whether the σ_i^2 are constant or not. If v_i denotes $\lambda_i^2 s_i^2 / n_i$, an approximate number of $d.f.$ are assigned to V by the rule (25):

$$d.f. = (\Sigma v_i)^2 / \Sigma \{v_i^2 / (n_i - 1)\}$$

When the n_i are all equal, this becomes $d.f. = (n - 1)(\Sigma v_i)^2 / \Sigma v_i^2$. For a test of significance we take $t = \Sigma \lambda_i \bar{X}_i / \sqrt{V}$, with this number of $d.f.$

To obtain an unbiased estimate of the error variance of $L = \Sigma \lambda_i \bar{X}_i$ in a two-way classification, calculate the comparison $L_j = \Sigma \lambda_i X_{ij}$ separately in every block, ($j = 1, 2, \dots, b$). The average of the b values L_j is, of course, L . The standard error of L is $\sqrt{\{\Sigma (L_j - L)^2 / b(b - 1)\}}$, with $(b - 1) d.f.$, which will be scanty if b is small.

If the trouble is caused by a few treatments whose means are substantially different from the rest, a satisfactory remedy is to omit these treatments from the main analysis, since conclusions about them are clear on inspection. With a one-way or two-way classification, the remaining treatments are analyzed in the usual way. The analysis of a Latin square with one omitted treatment is described in (17), and with two omitted treatments in (18).

11.14—Non-normality. Variance-stabilizing transformations. In the standard classifications, skewness in the distribution of errors tends to produce too many significant results in F - and t -tests. In addition, there is a loss of efficiency in the analysis, because when errors are non-normal the mean of the observed values for a treatment is, in general, not the most accurate estimate of the corresponding population mean for that treatment. If the mathematical form of the frequency distribution of the errors were known, a more efficient analysis could be developed. This approach is seldom attempted in practice, probably because the exact distribution of non-normal errors is rarely known and the more sophisticated analysis would be complicated.

With data in which the effects of the fixed factors are modest, there is some evidence that non-normality does not distort the conclusions too seriously. However, one feature of non-normal distributions is that the variance is often related to the mean. In the Poisson distribution, the variance equals the mean. For a binomial proportion with mean p , the variance is $p(1 - p)/n$. Thus, if treatment or replication effects are large, we expect unequal variances, with consequences similar to those discussed in the preceding section.

If σ_x^2 is a known function of the mean μ of X , say $\sigma_x^2 = \phi(\mu)$, a transformation of the data that makes the variance almost independent of the mean is obtained by an argument based on calculus. Let the transformation be $Y = f(X)$, and let $f'(X)$ denote the derivative of $f(X)$ with respect to X . By a one-term Taylor expansion

$$Y \doteq f(\mu) + f'(\mu)(X - \mu)$$

To this order of approximation, the mean value $E(Y)$ of Y is $f(\mu)$, since $E(X - \mu) = 0$. With the same approximation, the variance of Y is

$$E\{Y - f(\mu)\}^2 \doteq \{f'(\mu)\}^2 E(X - \mu)^2 = \{f'(\mu)\}^2 \sigma_x^2 = \{f'(\mu)\}^2 \phi(\mu)$$

Hence, to make the variance of Y independent of μ , we choose $f(\mu)$ so that the term on the extreme right above is a constant. This makes $f(\mu)$ the indefinite integral of $d\mu/\sqrt{\phi(\mu)}$. For the Poisson distribution, this gives $f(\mu) = \sqrt{\mu}$, i.e., $Y = \sqrt{X}$. For the binomial, the method gives $Y = \arcsin \sqrt{p}$, that is, Y is the angle whose sine is \sqrt{p} . When $f(X)$ has been chosen in this way, the value of the constant variance on the transformed scale is obtained by finding $\{f'(\mu)\}^2 \phi(\mu)$. For the Poisson, with $\phi(\mu) = \mu$, $f(\mu) = \sqrt{\mu}$, we have $f'(\mu) = 1/2\sqrt{\mu}$, so that $\{f'(\mu)\}^2 \phi(\mu) = \frac{1}{4}$. The variance on the transformed scale is $\frac{1}{4}$.

11.15—Square root transformation for counts. Counts of rare events, such as numbers of defects or of accidents, tend to be distributed approximately in Poisson fashion. A transformation to \sqrt{X} is often effective: the variance on the square root scale will be close to 0.25. If some counts are small, $\sqrt{X+1}$ or $\sqrt{X} + \sqrt{X+1}$, (19), stabilizes the variance more effectively.

TABLE 11 15 1
NUMBER OF POPPY PLANTS IN OATS
(Plants per 3 3/4 square feet)

Block	Treatment				
	A	B	C	D	E
1	438	538	77	17	18
2	442	422	61	31	26
3	319	377	157	87	77
4	380	315	52	16	20
Mean	395	413	87	38	35
Range	123	223	105	71	59

The square root transformation can also be used with counts in which it appears that the variance of X is *proportional* to the mean of X , that is, $\sigma_x^2 = k\bar{X}$. For a Poisson distribution of errors, $k = 1$, but we often find k larger than 1, indicating that the distribution of errors has a variance greater than that of the Poisson.

An example is the record of poppy plants in oats (20) shown in table 11 15 1, where the numbers are large. The differing ranges lead to a suspicion of heterogeneous variance. If the error mean square were calculated, it would be too large for testing differences among C , D , E and too small for A and B .

In table 11 15 2 the square roots of the numbers are recorded and analyzed. The ranges in the several treatments are now similar. That there are differences among treatments is obvious, it is unnecessary to compute F . The 5% LSD value is 3.09, suggesting that D and E are superior to C while, of course, the C , D , E group is much superior to A and B in reducing the numbers of undesired poppies.

TABLE 11 15 2
SQUARE ROOTS OF THE POPPY NUMBERS IN TABLE 11 15 1

Block	A	B	C	D	E
1	20.9	23.2	8.8	4.1	4.2
2	21.0	20.5	7.8	5.6	5.1
3	17.9	19.4	12.5	9.3	8.8
4	19.5	17.7	7.2	4.0	4.5
Mean	19.8	20.2	9.1	5.8	5.6
Range	3.1	5.5	5.3	5.3	4.6
Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square		
Blocks	3	22.65			
Treatments	4	865.44	216.36		
Error	12	48.69	4.06		

The means in the square root scale are reconverted to the original scale by squaring. This gives $(19.8)^2 = 392$ plants for *A*, $(20.2)^2 = 408$ plants for *B*, and so on. These values are slightly lower than the original means, 395 for *A*, 413 for *B*, etc., because the mean of a set of square roots is less than the square root of the original mean. As a rough correction for this discrepancy, add the Error mean square in the square root analysis to each reconverted mean. In this example we add 4.06, rounded to 4, giving 396 for *A*, and so on.

A transformation like the square root affects both the shape of the frequency distribution of the errors and the meaning of additivity. If treatment and block effects are additive in the original scale, they will not be additive in the square root scale, and vice versa. However, unless treatment and block effects are both large, effects that are additive in one scale will be approximately so in the other, since the square root transformation is a mild one.

EXAMPLE 11.15.1—The numbers of wireworms counted in the plots of a Latin square (21) following soil fumigations in the previous year were

Rows	Columns				
	1	2	3	4	5
1	<i>P</i> 3	<i>O</i> 2	<i>N</i> 5	<i>K</i> 1	<i>M</i> 4
2	<i>M</i> 6	<i>K</i> 0	<i>O</i> 6	<i>N</i> 4	<i>P</i> 4
3	<i>O</i> 4	<i>M</i> 9	<i>K</i> 1	<i>P</i> 6	<i>N</i> 5
4	<i>N</i> 17	<i>P</i> 8	<i>M</i> 8	<i>O</i> 9	<i>K</i> 0
5	<i>K</i> 4	<i>N</i> 4	<i>P</i> 2	<i>M</i> 4	<i>O</i> 8

Since these are such small numbers, transform to $\sqrt{(X+1)}$. The first number 3 becomes $\sqrt{3+1} = 2$, etc.

Analyze the variance. Ans. Mean square for Treatments 1.4457 for Error 0.3259.

EXAMPLE 11.15.2—Calculate the Studentized Range $D = 1.06$ and show that *K* gave significantly fewer wireworms than *M*, *N*, and *O*.

EXAMPLE 11.15.3—Estimate the average numbers of wireworms per plot for the several treatments. Ans. (with no bias correction) *K*, 0.99, *M*, 6.08, *N*, 6.40, *O*, 5.55, *P*, 4.38. To make the bias correction add 0.33 giving *K* = 1.32, *M* = 6.41, etc.

EXAMPLE 11.15.4—If the error variance of Y in the original scale is k times the mean of Y and if effects are additive in the square root scale, it can be shown that the true error variance in the square root scale is approximately $k/4$. Thus the value of k can be estimated from the analysis in the square root scale. If k is close to 1 this suggests that the distribution of errors in the original scale may be close to the Poisson distribution. In example 11.15.1 k is about $4(0.3259) = 1.3$ suggesting that most of the variance in the original scale is of the Poisson type. With the poppy plants (table 11.15.2) k is about 16 indicating a variance much greater than the Poisson.

11.16—Arcsin transformation for proportions. This transformation also called the *angular* transformation, was developed for binomial proportions. If a_{ij} successes out of n are obtained in the j th replicate of the i th treatment, the proportion $\hat{p}_{ij} = a_{ij}/n$ has variance $p_{ij}(1 - p_{ij})/n$. By means of table A.16, due to C. I. Bliss, we replace \hat{p}_{ij} by the angle whose

sine is $\sqrt{\hat{p}_{ij}}$. In the angular scale, proportions near 0 or 1 are spread out so as to increase their variance. If all the error variance is binomial, the error variance in the angular scale is about $821/n$. The transformation does not remove inequalities in variance arising from differing values of n . If the n 's vary widely, a weighted analysis in the angular scale is advisable.

With $n < 50$, a zero proportion should be counted as $1/4n$ before transforming to angles, and a 100% proportion as $(n - 1/4)/n$. This empirical device, suggested by Bartlett (22), improves the equality of variance in the angles. A more accurate transformation for small n has been tabulated by Mosteller and Youtz (19).

Angles may also be used with proportions that are subject to other sources of variation in addition to the binomial, if it is thought that the variance of \hat{p}_{ij} is some multiple of $p_{ij}(1 - p_{ij})$. Since, however, this product varies little for p_{ij} lying between 30% and 70%, the angular transformation is scarcely needed if nearly all the observed \hat{p}_{ij} lie in this range. In fact, this transformation is unlikely to produce a noticeable change in the conclusions unless the \hat{p}_{ij} range from near zero to 30% and beyond (or from below 70% to 100%).

Table 11.16.1, taken from a larger randomized blocks experiment (23), shows the percentages of unsalable ears of corn, the treatments being a control, *A*, and three mechanical methods of protecting against damage

TABLE 11.16.1
PERCENTAGE OF UNSALABLE EARS OF CORN

Treatments	Block							
	1	2	3	4	5	6		
<i>A</i>	42.4	34.3	24.1	39.5	55.5	49.1		
<i>B</i>	33.3	33.3	5.0	26.3	30.2	28.6		
<i>C</i>	8.5	21.9	6.2	16.0	13.5	15.4		
<i>D</i>	16.6	19.3	16.6	2.1	11.1	11.1		
	Angle = $\text{Arcsin } \sqrt{\text{Proportion}}$						Mean	%
<i>A</i>	40.6	35.8	29.4	38.9	48.2	44.5	39.6	40.6
<i>B</i>	35.2	35.2	12.9	30.9	33.3	32.3	29.9	24.9
<i>C</i>	17.0	27.9	14.4	23.6	21.6	23.1	21.3	13.2
<i>D</i>	24.0	26.1	24.0	8.3	19.5	19.5	20.2	11.9

Analysis of Variance in Angles

	Degrees of Freedom	Sum of Squares	Mean Square
Blocks	5	359.8	
Treatments	3	1,458.5	486.2
Error	15	546.1	36.4
Total	23	2,364.4	

by corn earworm larvae. The value of n , about 36, was not constant, but its variations were fairly small and are ignored. Note that the per cents range from 2.1% to 55.5%.

In the analysis of variance of the angles (table 11.16.1), the Error mean square was 36.4. Since $821/n = 821/36 = 22.8$, some variation in excess of the binomial may be present. The F -value for treatments is large. The 5% LSD for comparing two treatments is 7.4. B , C , and D were all superior to the control A , while C and D were superior to B . The angle means are retranslated to per cents at the right of the table.

11.17—The logarithmic transformation. Logarithms are used to stabilize the variance if the standard deviation in the original scale varies directly as the mean; in other words, if the coefficient of variation is constant. There are mathematical reasons why this type of relation between standard deviation and mean is likely to be found when the effects are *proportional* rather than *additive*; for example, when treatment 2 gives results consistently 23% higher than treatment 1 rather than results higher by, say, 18 units. In this situation the log transformation may bring about both additivity of effects and equality of variance. If some 0 values of X occur, $\log(X + 1)$ is often used.

TABLE 11.17.1
ESTIMATED NUMBERS OF FOUR KINDS OF PLANKTON (I . . . IV) CAUGHT IN SIX HAULS
WITH EACH OF TWO NETS

Haul	Estimated Numbers				Logarithms			
	I	II	III	IV	I	II	III	IV
1	895	1,520	43,300	11,000	2.95	3.18	4.64	4.04
2	540	1,610	32,800	8,600	2.73	3.21	4.52	3.93
3	1,020	1,900	28,800	8,260	3.01	3.28	4.46	3.92
4	470	1,350	34,600	9,830	2.67	3.13	4.54	3.99
5	428	980	27,800	7,600	2.63	2.99	4.44	3.88
6	620	1,710	32,800	9,650	2.79	3.23	4.52	3.98
7	760	1,930	28,100	8,900	2.88	3.29	4.45	3.95
8	537	1,960	18,900	6,060	2.73	3.29	4.28	3.78
9	845	1,840	31,400	10,200	2.93	3.26	4.50	4.01
10	1,050	2,410	39,500	15,500	3.02	3.38	4.60	4.19
11	387	1,520	29,000	9,250	2.59	3.18	4.46	3.97
12	497	1,685	22,300	7,900	2.70	3.23	4.35	3.90
Mean	671	1,701	30,775	9,396	2.802	3.221	4.480	3.962
Range	663	1,480	24,400	9,440	0.43	0.39	0.36	0.41

Analysis of Variance of Logarithms

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Kind of plankton	3	20.2070	6.7357
Haul	11	0.3387	0.0308
Discrepance	33	0.2300	0.0070

The plankton catches (24) of table 11.17.1 yielded nicely to the log transformation. The original ranges and means for the four kinds of plankton were nearly proportional, the ratios of range to mean being 0.99, 0.87, 0.79, and 1.00. After transformation the ranges were almost equal and uncorrelated with the means.

Transforming back, the estimated mean numbers caught for the four kinds of plankton are antilog $2.802 = 634$; 1,663; 30,200; and 9,162. These are *geometric means*.

The means of the logs will be found to differ significantly for all four kinds of plankton. The standard deviation of the logarithms is $\sqrt{0.0070} = 0.084$, and the antilogarithm of this number is 1.21. Quoting Winsor and Clark (page 5), "Now a deviation of 0.084 in the logarithms of the catch means that the catch has been multiplied (or divided) by 1.21. Hence we may say that one standard deviation in the logarithm corresponds to a percentage standard deviation, or coefficient of variation, of 21% in the catch."

EXAMPLE 11.17.1—The following data were abstracted from an experiment (27) which was more complicated in design. Each entry is the geometric mean of insect catches by a trap in three successive nights, one night at each of three locations. Three types of trap are compared over five three-night periods. The insects are macrolepidoptera at Rothamsted Experimental Station:

Trap	3-Night Periods. August, 1950				
	16-18	19-21	22-24	25-27	28-30
1	19.1	23.4	29.5	23.4	16.6
2	50.1	166.1	223.9	58.9	64.6
3	123.0	407.4	398.1	229.1	251.2

Williams found the log transformation effective in analyzing highly variable data like these. Transform to logarithms and analyze their variance. Ans. Mean square for traps = 1.4455; for error, 0.0172.

Show that all differences between trap means are significant and that the geometric means for traps are 21.9, 93.3, and 257.0 insects.

11.18—Non-additivity. Suppose that in a two-way classification, with 2 rows and 2 columns, the effects of rows and columns are proportional or multiplicative instead of additive. In each row, column *B* exceeds column *A* by a fixed percentage, while in each column, row 2 exceeds row 1 by a fixed percentage. Consider column percentages of 20% and 100% and row percentages of 10% and 50%. These together provide four combinations. Taking the observation in column *A*, row 1, as 1.0, the other observations are shown in table 11.18.1 for the four cases.

Thus, in case 1, the value of 1.32 for *B* in row 2 is 1.1×1.2 . Since no experimental error has been added, the error mean square in a correct analysis should be zero. The correct procedure is to transform the data to logs before analysis. In logs the effects become additive, and the error mean square is zero. From the analysis in logs, we learn that *B* exceeds *A* by exactly 20% in cases 1 and 2, and by exactly 100% in cases 3 and 4.

TABLE 11.18.1
HYPOTHETICAL DATA FOR FOUR CASES WITH MULTIPLICATIVE EFFECTS

Row	Case 1 C 20% R 10%		Case 2 C 20% R 50%		Case 3 C 100% R 10%		Case 4 C 100% R 50%	
	A	B	A	B	A	B	A	B
1	1.0	1.2	1.0	1.2	1.0	2.0	1.0	2.0
2	1.1	1.32	1.5	1.8	1.1	2.2	1.5	3.0
Means	1.05	1.26	1.25	1.50	1.05	2.10	1.25	2.50
s		0.01		0.05		0.05		0.25
s/\bar{X}		0.9%		3.6%		3.2%		13.3%

If the usual analysis of variance is carried out in the *original* scale, the standard error s per observation (with 1 *d.f.*) is shown under each case. With 2 replications, s is also the *s.e.* of the difference $\bar{B} - \bar{A}$. Consequently, in case 1 we would conclude from this analysis that $\bar{B} - \bar{A}$ is 0.21 with a standard error of ± 0.01 . In case 4 we conclude that $\bar{B} - \bar{A} = 1.25 \pm 0.25$. The standard errors, ± 0.01 and ± 0.25 , are entirely a result of the fact that we used the wrong model for analysis. In a real situation where experimental errors are also present, this variance s^2 due to non-additivity is added to the ordinary experimental error variance σ^2 .

To generalize, the analysis in the original scale has two defects. It fails to discover the simple proportional nature of the relationship between row and column effects. It also suffers a loss of precision, since the error variance is inflated by the component due to non-additivity. If row and column effects are both small, these deficiencies are usually not serious. In case 1, for example, the standard error s due to non-additivity only is 0.9% of the mean. If the ordinary standard error σ were 5% of the mean (a low value for most data), the non-additivity would increase this only to $\sqrt{25.81}$ or 5.1%. The loss of precision from non-additivity is greater in cases 2 and 3 and jumps markedly in case 4 in which both row and column effects are large.

11.19—Tukey's test of additivity. This is useful in a variety of ways: (i) to help decide if a transformation is necessary; (ii) to suggest a suitable transformation; (iii) to learn if a transformation has been successful in producing additivity (28, 29).

The test is related to transformations of the form $Y = X^p$, in which X is the original scale, and we are seeking a power p of X such that effects are additive in the scale of $Y = X^p$. Thus, $p = 1/2$ represents the square root transformation and $p = -1$ a *reciprocal* transformation, analyzing $1/X$ instead of X . The value $p = 0$ is interpreted as a log transformation, because the variable X^p behaves like $\log X$ when p is small.

The rationale of the test can be indicated by means of calculus. For

the two-way classification, if effects are exactly additive in the scale of Y , we have,

$$\begin{aligned} Y_{ij} &= \bar{Y}_{..} + (\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{.j} - \bar{Y}_{..}) \\ &= \bar{Y}_{..} [1 + \{(\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{.j} - \bar{Y}_{..})\} / \bar{Y}_{..}] \end{aligned}$$

We suppose that row and column effects are small relative to the mean. This implies that $\alpha_i = (\bar{Y}_{i.} - \bar{Y}_{..}) / \bar{Y}_{..}$ and $\beta_j = (\bar{Y}_{.j} - \bar{Y}_{..}) / \bar{Y}_{..}$ are both small.

Write $X_{ij} = Y_{ij}^{1/p}$ and expand in the usual Taylor's series. This gives

$$\begin{aligned} X_{ij} &= \bar{Y}_{..}^{1/p} [1 + \alpha_i + \beta_j]^{1/p} \\ &= \bar{Y}_{..}^{1/p} \left[1 + \frac{1}{p} (\alpha_i + \beta_j) + \frac{1}{p} \frac{(1-p)}{p} \frac{1}{2} (\alpha_i^2 + 2\alpha_i\beta_j + \beta_j^2) + \dots \right] \end{aligned}$$

Now, in the X scale the terms in α_i, α_i^2 represent row effects and the terms in β_j, β_j^2 represent column effects that are added together in the above expression. These terms are therefore still additive in the X scale. The first non-additive term is the one in $\alpha_i\beta_j$. Written in full, this term is

$$\bar{Y}_{..}^{1/p} (1-p) (\bar{Y}_{i.} - \bar{Y}_{..}) (\bar{Y}_{.j} - \bar{Y}_{..}) / p^2 \bar{Y}_{..}^2 \quad (11.19.1)$$

For our purpose we need to write this expression in terms of X rather than Y . By new single-term Taylor expansions we have, since $Y = X^p$,

$$\bar{Y}_{i.} - \bar{Y}_{..} \doteq p \bar{X}_{..}^{p-1} (\bar{X}_{i.} - \bar{X}_{..}): \quad \bar{Y}_{.j} - \bar{Y}_{..} \doteq p \bar{X}_{..}^{p-1} (\bar{X}_{.j} - \bar{X}_{..})$$

Substitution into (11.19.1) gives for the first non-additive term in X_{ij} ,

$$(1-p) \bar{Y}_{..}^{1/p} (\bar{X}_{i.} - \bar{X}_{..}) (\bar{X}_{.j} - \bar{X}_{..}) \bar{X}_{..}^{2p-2} / \bar{Y}_{..}^2$$

Using $\bar{Y}_{..} \doteq \bar{X}_{..}^p$, this term may be expressed approximately as

$$\frac{(1-p)}{\bar{X}_{..}} (\bar{X}_{i.} - \bar{X}_{..}) (\bar{X}_{.j} - \bar{X}_{..}) \quad (11.19.2)$$

Since this term represents a non-additive effect of rows and columns, it will appear in the residual of X_{ij} when an additive model is fitted in the X scale. The conclusions from this rough argument are as follows:

1. If this type of non-additivity is present in X , and \hat{X}_{ij} is the fitted value given by the additive model, the residual $X_{ij} - \hat{X}_{ij}$ has a linear regression on the variate $(\bar{X}_{i.} - \bar{X}_{..})(\bar{X}_{.j} - \bar{X}_{..})$.

2. The regression coefficient B is an estimate of $(1-p)/\bar{X}_{..}$. Thus, the power p to which X must be raised to produce additivity is estimated by $(1 - B\bar{X}_{..})$. Commenting on this result, Anscombe and Tukey (13) state (their k is our $B/2$): "It is important to emphasize that the available data rarely define the 'correct' value of p with any precision. Repeating the analysis and calculation of k for each of a number of values of p may show the range of values of p clearly compatible with the observations, but experience and subject-matter insight are important in choosing a p for final analysis."

3. Tukey's test is a test of the null hypothesis that the population value of B is zero. A convenient way of computing B and making the test is illustrated by the data in table 11.19.1. The data are average insect catches in three traps over five periods. The same data were presented in example 11.17.1 as an exercise on the log transformation. We now consider the additivity of trap and period effects in the original scale. The steps are as follows (see table 11.19.1 for calculations):

TABLE 11.19.1
MACROLEPIDOPTERA CATCHES BY THREE TRAPS IN FIVE PERIODS
(Calculations for test of additivity)

Period	1	Trap 2	3	Sum X_i	Mean \bar{X}_i	d_i	$n_i = \sum X_{ij}d_j$
1	19.1	50.1	123.0	192.2	64.1	-74.9	14,025
2	23.4	166.1	407.4	596.9	199.0	+60.0	51,096
3	29.5	223.9	398.1	651.5	217.2	+78.2	47,543
4	23.4	58.9	229.1	311.4	103.8	-35.2	28,444
5	16.6	64.6	251.2	332.4	110.8	-28.1	32,243
Sum $X_{.j}$	112.0	563.6	1408.8	2084.4			$\sum n_i = 173,351$
Mean $\bar{X}_{.j}$	22.4	112.7	281.8		139.0		
d_j	-116.6	-26.2	+142.8			0.0	

- (i) Find $d_i = X_i - \bar{X}_{..}$ and $d_j = \bar{X}_{.j} - \bar{X}_{..}$, both adding exactly to zero
(ii) $w_1 = (19.1)(-116.6) + (50.1)(-26.2) + (123.0)(+142.8) = 14,025$
 $w_5 = (16.6)(-116.6) + (64.6)(-26.2) + (251.2)(+142.8) = 32,243$
Check: $173,351 = (112.0)(-116.6) + (563.6)(-26.2) + (1408.8)(+142.8)$
 $N = \sum w_i d_i = (14,025)(-74.9) + \dots + (32,243)(-28.1) = 3.8259 \times 10^6$
(iii) $\sum d_i^2 = (-74.9)^2 + \dots + (-28.1)^2 = 17,354$
 $\sum d_j^2 = (-116.6)^2 + \dots + (+142.8)^2 = 34,674$
 $D = (\sum d_i^2)(\sum d_j^2) = (17,354)(34,674) = 601.7 \times 10^6$
(iv) SS for non-additivity = $\frac{N^2}{D} = \frac{(3.8259)^2(10^{12})}{(601.7)(10^6)} = 24,327$

(i) Calculate $d_i = \bar{X}_i - \bar{X}_{..}$ and $d_j = \bar{X}_{.j} - \bar{X}_{..}$, rounding if necessary so that both sets add *exactly* to zero.

(ii) Compute $w_i = \sum_j X_{ij}d_j$ and record them in the extreme right column: Then find

$$N = \sum_i w_i d_i = \sum \sum X_{ij} d_i d_j$$

N is the numerator of B

(iii) The denominator D of B is $(\sum d_i^2)(\sum d_j^2)$. Thus, $B = N/D$.

(iv) The contribution of non-additivity to the error sum of squares of χ is N^2/D , with 1 d.f. This is tested by an F -test against the remainder

TABLE 11.19.2
ANALYSIS OF VARIANCE AND TEST OF ADDITIVITY

	Degrees of Freedom	Sum of Squares	Mean Square
Periods	4	52,066	
Traps	2	173,333	
Error	8	30,607	
Non-additivity	1	24,327	24,327
Remainder	7	6,280	897

$$F = 24,327/897 = 27.1, \text{ d.f.} = 1, 7. P < 0.01$$

of the Error *S.S.*, which has $\{(r-1)(c-1)-1\}$ *d.f.* The test is made in table 11.19.2.

The hypothesis of additivity is untenable. What type of transformation is suggested by this test?

$$B = \frac{N}{D} = \frac{3.8259}{601.7} = 0.006358$$

$$\hat{p} = 1 - B\bar{X}_{..} = 1 - (0.006358)(139.0) = 1 - 0.88 = 0.12.$$

The test suggests a one-tenth power of X . This behaves very much like $\log X$.

11.20—Non-additivity in a Latin square. If the mathematical analysis of the previous section is carried out for a Latin square, the first non-additive term, corresponding to equation 11.19.2, is, as might be guessed,

$$\frac{(1-p)}{\bar{X}_{...}} \{(\bar{X}_{i..} - \bar{X}_{...})(\bar{X}_{.j.} - \bar{X}_{...}) + (\bar{X}_{i..} - \bar{X}_{...})(\bar{X}_{..k} - \bar{X}_{...}) \\ + (\bar{X}_{.j.} - \bar{X}_{...})(\bar{X}_{..k} - \bar{X}_{...})\}$$

Consequently, the test for additivity is carried out by finding the regression of $(X_{ijk} - \hat{X}_{ijk})$ on the variate in $\{ \}$ above, as illustrated in (28). Note that D is the error sum of squares of the $\{ \}$ variable.

We shall, instead, illustrate an alternative method of doing the computations, due to Tukey (29), that generalizes to other classifications. Table 11.20.1 comes from an experiment on monkeys (30), the raw data being the number of responses to auditory or visual stimuli administered under five conditions (A, \dots, E). Each pair of monkeys received one type of stimulus per week, the order from week to week being determined by the randomized columns of the Latin square.

It was discovered that the standard deviation of the number of responses was almost directly proportional to the mean, so the counts were transformed to logs. Each entry in the table is the mean of the log counts for the two members of a pair. Has additivity been attained?

TABLE 11.20.1
LOGS OF NUMBERS OF RESPONSES BY PAIRS OF MONKEYS UNDER FIVE STIMULI
(Test of additivity in a Latin square)

Pair	Week										$\bar{X}_{i..}$
	1	2	3	4	5						
1	B	1.99	D	2.25	C	2.18	A	2.18	E	2.51	2.222
\hat{X}_{ijk}		2.022		2.268		2.220		2.084		2.518	
d_{ijk}		-0.032		-0.018		-0.040		0.098*		-0.008	
U_{ijk}		37		3		0		17		92	
2	D	2.00	B	1.85	A	1.79	E	2.14	C	2.31	2.018
		1.950		1.932		1.852		2.152		2.206	
		0.052*		-0.082		-0.062		-0.012		0.104	
		70		80		132		4		0	
3	C	2.17	A	2.10	E	2.34	B	2.20	D	2.40	2.242
		2.132		2.082		2.348		2.178		2.472	
		0.038		0.018		-0.006*		0.022		-0.072	
		7		18		18		1		66	
4	E	2.41	C	2.47	B	2.44	D	2.53	A	2.44	2.458
		2.456		2.462		2.366		2.526		2.482	
		-0.046		0.010*		0.074		0.004		-0.042	
		58		61		23		97		71	
5	A	1.85	E	2.32	D	2.21	C	2.05	B	2.25	2.136
		1.862		2.248		2.176		2.162		2.234	
		-0.012		0.072		0.034		-0.112		0.018*	
		125		1		2		3		0	
$\bar{X}_{.j.}$		2.084		2.198		2.192		2.220		2.382	2.215
$\bar{X}_{..k}$		A		B		C		D		E	
		2.072		2.146		2.236		2.278		2.344	

* Denotes deviations that were adjusted in order to make the deviations add to zero over every row, column, and treatment.

The steps follow.

1. Find the row, column, and treatment means, as shown, and the fitted values \hat{X}_{ijk} by the additive model

$$\hat{X}_{ijk} = \bar{X}_{i..} + \bar{X}_{.j.} + \bar{X}_{..k} - 2\bar{X}...$$

For E in row 2, column 4,

$$\hat{X}_{245} = 2.018 + 2.220 + 2.344 - 2(2.215) = 2.152$$

2. Find the residuals $d_{ijk} = X_{ijk} - \hat{X}_{ijk}$ as shown, adjusting if neces-

sary so that the sums are zero over every row, column, and treatment. Values that were adjusted are denoted by an * in table 11.20.1.

3. Construct the 25 values of a variate $U_{ijk} = c_1(\hat{X}_{ijk} - c_2)^2$, where c_1 and c_2 are any two convenient constants. We took $c_2 = \bar{X} \dots = 2.215$, which is often suitable, and $c_1 = 1000$, so that the U 's are mostly between 0 and 100. For B in row 1, column 1,

$$U_{112} = 1000(2.022 - 2.215)^2 = 37$$

4. Calculate the regression coefficient of the d_{ijk} on the residuals of the U_{ijk} . The numerator is

$$N = \sum d_{ijk} U_{ijk} = (-0.032)(37) + \dots + (0.018)(0) = -20.356$$

The denominator D is the error sum of squares of the U_{ijk} . This is found by performing the ordinary Latin square analysis of the U_{ijk} . The value of D is 22,330.

5. To perform the test for additivity, find the $S.S.$, 0.0731, of the d_{ijk} , which equals the error $S.S.$ of the X_{ijk} . The contribution due to non-additivity is $N^2/D = (-20.356)^2/22,330 = 0.0186$. Finally, compare the mean square for Non-additivity with the Remainder mean square.

	Degrees of Freedom	Sum of Squares	Mean Square	F
Error S S	12	0.0731		
Non-additivity	1	0.0186	0.0186	3.76 ($P = 0.08$)
Remainder	11	0.0545	0.00495	

The value of P is 0.08—a little low, though short of the 5% level. Since the interpretations are not critical (examples 11.20.4, 11.20.5), the presence of slight non-additivity should not affect them.

The above procedure applies also in more complex classifications. Note that if we expand the quadratic $c_1(\hat{X}_{ijk} - \bar{X} \dots)^2$, the coefficient of terms like $(\bar{X}_{i.} - \bar{X} \dots)(\bar{X}_{.j} - \bar{X} \dots)$ is $2c_2$. Hence the regression coefficient B of the previous section is $B = 2c_1 N/D$. If a power transformation is needed, the suggested power is as before $p = 1 - B\bar{X} \dots$.

EXAMPLE 11 20 1—The following data are the number of lesions on eight pairs of half leaves inoculated with two strengths of tobacco virus (from table 4 3 1)

Treatments	Replications							
	1	2	3	4	5	6	7	8
1	31	20	18	17	9	8	10	7
2	18	17	14	11	10	7	5	6

Test for additivity by the method of section 11.19. Ans.:

	Degrees of Freedom	Sum of Squares	Mean Square	F
Error	7	65		
Non-additivity	1	38	38	8.4
Remainder	6	27	4.5	

F is significant at the 5% level. The non-additivity may be due to anomalous behavior of the 31, 18 pair.

EXAMPLE 11.20.2—Apply $\sqrt{(X+1)}$ to the virus data. While F now becomes non-significant, the pair (31, 18) still appears unusual.

EXAMPLE 11.20.3—The data in example 11.2.1, regarded as a 3×3 two-way classification, provide another simple example of Tukey's test. Ans. For non-additivity, $F = 5.66$.

EXAMPLE 11.20.4—Analyze the variance of the logarithms of the monkey responses. You will get,

	Degrees of Freedom	Sum of Squares	Mean Square	F
Monkey Pairs	4	0.5244	0.1311	
Weeks	4	0.2294	0.0574	
Stimuli	4	0.2313	0.0578	9.6
Error	12	0.0725	0.00604	

EXAMPLE 11.20.5—Test all differences among the means in table 11.20.1, using the LSD method. Ans. $E > A, B, C$; $D > A, B$; $C > A$.

EXAMPLE 11.20.6—Calculate the sum of squares due to the regression of log response on weeks. It is convenient to code the weeks as $X = -2, -1, 0, 1, 2$. Then, taking the weekly means as Y , $\Sigma xy = 0.618$ and $(\Sigma xy)^2 / \Sigma x^2 = 0.03819$. On the per item basis, the sum of squares due to regression is $5(0.03819) = 0.1910$. The line for Weeks in example 11.20.4 may now be separated into two parts:

Linear Regression	1	0.1910	0.1910
Deviations from Regression	3	0.0384	0.0128

Comparing the mean squares with error, it is seen that deviations are not significant, most of the sum of squares for Weeks being due to the regression.

REFERENCES

1. R. H. PORTER. *Cooperative Soybean Seed Treatment Trials*, Iowa State College Seed Laboratory (1936).
2. S. P. MONSELISE. *Palestine J. Botany*, 8:1 (1951).
3. W. G. COCHRAN and G. M. COX. *Experimental Designs*, 2nd ed., Wiley, New York (1957).
4. O. KEMPTHORNE. *Design and Analysis of Experiments*. Wiley, New York (1952).
5. R. A. FISHER. *The Design of Experiments*. Oliver and Boyd, Edinburgh (1935–1951).
6. H. C. FORSTER and A. J. VASEY. *J. Dept. of Agric.*, Victoria, Australia, 30–35 (1932).
7. R. A. FISHER and F. YATES. *Statistical Tables*. Oliver and Boyd, Edinburgh (1938–1953).
8. W. T. FENG and T. N. LIU. *J. Amer. Soc. Agron.*, 28:1 (1936).

338 Chapter 11: Two-Way Classifications

- 9 W. G. COCHRAN, K. M. AUTREY, and C. Y. CANNON. *J. Dairy Sci.*, 24:937 (1941).
- 10 M. HEALY and M. WESTMACOTT. *Applied Statistics*, 5:203 (1956).
- 11 H. SCHEFFE. *The Analysis of Variance*. Wiley, New York (1959).
- 12 F. J. ANSCOMBE. *Technometrics*, 2:123 (1960).
- 13 F. J. ANSCOMBE and J. W. TUKEY. *Technometrics*, 5:141 (1963).
- 14 W. G. COCHRAN. *Biometrics*, 3:33 (1947).
- 15 W. T. FEDERER and C. S. SCHLOTTFELDT. *Biometrics*, 10:282-90 (1954).
- 16 A. D. OUTHWAITE and A. RUTHERFORD. *Biometrics*, 11:431 (1955).
- 17 F. YATES. *J. Agric. Sci.*, 26:301 (1936).
- 18 F. YATES and R. W. HALE. *J. R. Statist. Soc. Suppl.*, 6:67 (1939).
- 19 F. MOSTELLER and C. YOUTZ. *Biometrika*, 48:433 (1961).
- 20 M. S. BARTLETT. *J. R. Statist. Soc. Suppl.*, 3:68 (1936).
- 21 W. G. COCHRAN. *Emp. J. Exp. Agric.*, 6:157 (1938).
- 22 M. S. BARTLETT. *Biometrics*, 3:39 (1947).
- 23 W. G. COCHRAN. *Ann. Math. Statist.*, 11:344 (1940).
- 24 C. P. WINSOR and G. L. CLARKE. Sears Foundation: *J. Marine Res.*, 3:1 (1940).
- 25 F. E. SATTERTHWAITE. *Biometrics Bull.*, 2:110 (1946).
- 26 F. YATES. *Emp. Jour. Exp. Agric.*, 1:129 (1933).
- 27 C. B. WILLIAMS. *Bul. Entomological Res.*, 42:513 (1951).
- 28 J. W. TUKEY. *Biometrics*, 5:232 (1949).
- 29 J. W. TUKEY. *Queries in Biometrics*, 11:111 (1955).
- 30 R. A. BUTLER. *J. Exp. Psych.*, 48:19 (1954).

Factorial experiments

12.1—Introduction. A common problem in research is investigating the effects of each of a number of variables or *factors* as they are called, on some response Y . Suppose a company in the food industry proposes to market a cake mix from which the housewife can make a cake by adding water and then baking. The company must decide on the best kind of flour and the correct amounts of fat, sugar, liquid (milk or water), eggs, baking powder, and flavoring, as well as on the best oven temperature and the proper baking time. These are nine factors, any one of which may affect the palatability and the keeping quality of the cake to a noticeable degree. Similarly, a research program designed to learn how to increase the yields of the principal cereal crop in a country is likely to try to measure the effects on yield of different amounts of nitrogen, phosphorus, and potassium when added as fertilizers to the soil. Problems of this type occur frequently in industry: with complex chemical processes there can be as many as 10 to 20 factors that may affect the final product.

In earlier times the advice was sometimes given to study one factor at a time, a separate experiment being devoted to each factor. Later, Fisher (1) pointed out that important advantages are gained by combining the study of several factors in the same *factorial experiment*. Factorial experimentation is highly efficient, because every observation supplies information about all the factors included in the experiment. Secondly, as we will see, factorial experimentation is a workmanlike method of investigating the relationships between the effects of different factors.

12.2—The single factor versus the factorial approach. To illustrate the difference between the “one factor at a time” approach and the factorial approach, consider an investigator who has two factors, A and B , to study. For simplicity, suppose that only two *levels* of each factor, say a_1, a_2 , and b_1, b_2 are to be compared. In a cake mix, a_1, a_2 might be two types of flour and b_1, b_2 two amounts of flavoring. Four replications are considered sufficient by the investigator.

In the single-factor approach, the first experiment is a comparison of a_1 with a_2 . The level of B is kept constant in the first experiment, but

the investigator must decide what this constant level is to be. We shall suppose that B is kept at b_1 : the choice made does not affect our argument. The two treatments in the first experiment may be denoted by the symbols a_1b_1 and a_2b_1 , replicated four times. The effect of A , that is, the mean difference $a_2b_1 - a_1b_1$, is estimated with a variance $2\sigma^2/4 = \sigma^2/2$.

The second experiment compares b_2 with b_1 . If a_2 performed better than a_1 in the first experiment, the investigator is likely to use a_2 as the constant level of A in the second experiment (again, this choice is not vital to the argument). Thus, the second experiment compares a_2b_1 with a_2b_2 in four replications, and estimates the effect of B with variance $\sigma^2/2$.

In the two single-factor experiments, 16 observations have been made, and the effects of A and B have each been estimated with variance $\sigma^2/2$.

But suppose that someone else, interested in these factors, hears that experiments on them have been done. He asks the investigator: In my work, I have to keep A at its lower level, a_1 . What effect does B have when A is at a_1 ? Obviously, the investigator cannot answer this question, since he measured the effect of B only when A was held at its higher level. Another person might ask: Is the effect of A the same at the two levels of B ? Once again, the investigator has no answer, since A was tested at only one level of B .

In the factorial experiment, the investigator compares all treatments that can be formed by combining the levels of the different factors. There are four such treatment combinations, a_1b_1 , a_2b_1 , a_1b_2 , a_2b_2 . Notice that each replication of this experiment supplies *two* estimates of the effect of A . The comparison $a_2b_2 - a_1b_2$ estimates the effect of A when B is held constant at its higher level, while the comparison $a_2b_1 - a_1b_1$ estimates the effect of A when B is held constant at its lower level. The average of these two estimates is called the *main effect* of A , the adjective *main* being a reminder that this is an average taken over the levels of the other factor. In terms of our definition of a comparison (section 10.7) the main effect of A may be expressed as

$$L_A = \frac{1}{2}(a_2b_2) + \frac{1}{2}(a_2b_1) - \frac{1}{2}(a_1b_2) - \frac{1}{2}(a_1b_1), \quad (12.2.1)$$

where (a_2b_2) denotes the yield given by the treatment combination a_2b_2 (or the average yield if the experiment has r replications), and so on. By Rule 10.7.1 the variance of L_A is

$$\frac{\sigma^2}{r} \left\{ \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right\} = \frac{\sigma^2}{r}$$

If the investigator uses 2 replications (8 observations), the main effect of A is estimated with a variance $\sigma^2/2$.

Now consider B . Each replication furnishes two estimates, $a_2b_2 - a_2b_1$, and $a_1b_2 - a_1b_1$, of the effect of B . The main effect of B is the comparison

$$L_B = \frac{1}{2}(a_2b_2) + \frac{1}{2}(a_1b_2) - \frac{1}{2}(a_2b_1) - \frac{1}{2}(a_1b_1) \quad (12.2.2)$$

With two replications of the factorial experiment (8 observations), L_B , like L_A , has variance $\sigma^2/2$.

Thus, the factorial experiment requires only 8 observations, as against 16 by the single-factor approach, to estimate the effects of A and B with the same variance $\sigma^2/2$. With 3 factors, the factorial experiment requires only 1/3 as many observations, with 4 factors only 1/4, and so on. These striking gains in efficiency occur because every observation, like (a_2b_1) , or $(a_1b_2c_2)$, or $(a_2b_1c_1d_2)$, is used in the estimate of the effect of every factor. In the single-factor approach, on the other hand, an observation supplies information only about the effect of one factor.

What about the relationship between the effects of the factors? The factorial experiment provides a separate estimate of the effects of A at each level of B , though these estimates are less precise than the main effect of A , their variance being σ^2 . The question: Is the effect of A the same at the two levels of B ?, can be examined by means of the comparison:

$$\{(a_2b_2) - (a_1b_2)\} - \{(a_2b_1) - (a_1b_1)\} \quad (12.2.3)$$

This expression measures the difference between the effect of A when B is at its higher level and the effect of A when B is at its lower level. If the question is: Does the level of A influence the effect of B ?, the relevant comparison is

$$\{(a_2b_2) - (a_2b_1)\} - \{(a_1b_2) - (a_1b_1)\} \quad (12.2.4)$$

Notice that (12.2.3) and (12.2.4) are identical. The expression is called the *AB two-factor interaction*. In this, the combinations (a_2b_2) and (a_1b_1) receive a + sign, the combinations (a_2b_1) and (a_1b_2) a - sign.

Because of its efficiency and comprehensiveness, factorial experimentation is extensively used in research programs, particularly in industry. One limitation is that a factorial experiment is usually larger and more complex than a single-factor experiment. The potentialities of factorial experimentation in clinical medicine have not been fully exploited, because it is usually difficult to find enough suitable patients to compare more than two or three treatment combinations.

In analyzing the results of a 2^2 factorial, the commonest procedure is to look first at the two main effects and the two-factor interaction. If the interaction seems absent, we need only report the main effects, with some assurance that each effect holds at either level of the other variate. A more compact notation for describing the treatment combinations is also standard. The presence of a letter a or b denotes one level of the factor in question, while the absence of the letter denotes the other level. Thus, a_2b_2 becomes ab , and a_1b_2 becomes b . The combination a_1b_1 is denoted by the symbol (1). In this notation, table 12.2.1 shows how to compute the main effects and the interaction from the treatment totals over r replications.

TABLE 12.2.1
CALCULATION OF MAIN EFFECTS AND INTERACTION IN A 2^2 FACTORIAL

Factorial Effect:	Multiplier for Treatment Total				Divisor to give Mean	Contribution to Treatments S.S.
	(1)	<i>a</i>	<i>b</i>	<i>ab</i>		
<i>A</i>	-1	1	-1	1	2 <i>r</i>	$[A]^2/4r$
<i>B</i>	-1	-1	1	1	2 <i>r</i>	$[B]^2/4r$
<i>AB</i>	1	-1	-1	1	2 <i>r</i>	$[AB]^2/4r$

Thus, the main effect of *A* is:

$$[A]/2r = [(ab) - (b) + (a) - (1)]/2r$$

The quantities $[A]$, $[B]$, $[AB]$ are called factorial effect totals. Use of the same divisor, $2r$, for the AB interaction mean is a common convention.

In the analysis of variance, the contribution of the main effect of *A* to the Treatments S.S. is $[A]^2/4r$, by Rule 11.6.1. Further, note that the three comparisons $[A]$, $[B]$ and $[AB]$ in table 12.2.1 are orthogonal. By Rule 11.6.4, the three contributions in the right-hand column of table 12.2.1 add up to the Treatments S.S.

EXAMPLE 12.2.1—Yates (2) pointed out that the concept of factorial experimentation can be applied to gain accuracy when weighing objects on a balance with two pans. Suppose that two objects are to be weighed and that in any weighing the balance has an error distributed about 0 with variance σ^2 . If the two objects are weighed separately, the balance estimates each weight with variance σ^2 . Instead, both objects are placed in one pan, giving an estimate y_1 of the sum of the weights. Then the objects are placed in different pans, giving an estimate y_2 of the difference between the weights. Show that the quantities $(y_1 + y_2)/2$ and $(y_1 - y_2)/2$ give estimates of the individual weights with variance $\sigma^2/2$.

EXAMPLE 12.2.2—If four objects are to be weighed, show how to conduct four weighings so that the weight of each object is estimated with variance $\sigma^2/4$. Hint: First weigh the sum of the objects, then refer to table 12.2.1.

12.3—Analysis of the 2^2 factorial experiment. The case where no interaction appears is illustrated by an experiment (3) on the fluorometric determination of the riboflavin content of dried collard leaves (table 12.3.1). The two factors were *A*, the size of sample (0.25 gm., 1.00 gm.) from which the determination was made, and *B*, the effect of the inclusion of a permanganate-peroxide clarification step in the determination. This was a randomized blocks design replicated on three successive days.

The usual analysis of variance into Replications, Treatments, and Error is computed. Then the factorial effect totals for *A*, *B*, and AB are calculated from the treatment totals, using the multipliers given in table 12.3.1. Their squares are divided by $4r$, or 12, to give the contributions to the Treatments S.S. The *P* value corresponding to the *F* ratio 13.02/8.18 for Interaction is about 0.25: we shall assume interaction absent. Consequently, attention can be concentrated on the main effects. The Permanganate step produced a large reduction in the estimated riboflavin concentration. The effect of Sample Size was not quite significant.

TABLE 12.3.1
APPARENT RIBOFLAVIN CONCENTRATION (MCG./GM.) IN COLLARD LEAVES

Replication	Without Permanganate		With Permanganate		Total	
	0.25 gm. Sample	1.00 gm. Sample	0.25 gm. Sample	1.00 gm. Sample		
1	39.5	38.6	27.2	24.6	129.9	
2	43.1	39.5	23.2	24.2	130.0	
3	45.2	33.0	24.8	22.2	125.2	
Total	127.8 (1)	111.1 <i>a</i>	75.2 <i>b</i>	71.0 <i>ab</i>	Factorial Effect Total	Factorial Effect Mean S.E.
Sample Size (<i>A</i>)	-1	1	-1	1	-20.9	- 3.5
Permanganate (<i>B</i>)	-1	-1	1	1	-92.7	-15.4
Interaction (<i>AB</i>)	1	-1	-1	1	12.5	2.1

} ± 1.65

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	<i>P</i>
Replications	2	3.76		
Treatments	(3)	(765.53)		
Sample size	1	$(-20.9)^2/12 = 36.40$	36.40	0.08
Permanganate	1	$(-92.7)^2/12 = 716.11$	716.11	<0.01
Interaction	1	$(12.5)^2/12 = 13.02$	13.02	0.25
Error	6	49.08	8.18	

Instead of subdividing the Treatments S.S. and making *F*-tests, one can proceed directly to compute the factorial effect means. These are obtained by dividing the effect totals by 2*r*, or 6, and are shown in table 12.3.1 beside the effect totals. The standard error of an effect mean is $\sqrt{s^2/r} = \sqrt{2.73} = 1.65$. The *t*-tests of the effect means are of course the same as the *F*-tests in the analysis of variance. Use of the effect means has the advantage of showing the magnitude and direction of the effects.

The principal conclusion from this experiment was that "In the fluorometric determination of riboflavin of the standard dried collard sample, the permanganate-hydrogen peroxide clarification step is essential. Without this step, the mean value is 39.8 mcg. per gram, while with it the more reasonable mean of 24.4 is obtained." These data are discussed further in example 12.4.1.

EXAMPLE 12.3.1—From table 12.3.1, calculate the means of the four treatment combinations. Then calculate the main effects of *A* and *B*, and verify that they are the same as the "Effect Means" shown in table 12.3.1. Verify also that the *AB* interaction, if calculated by equations (12.2.3) or (12.2.4), is twice the effect mean in table 12.3.1. As already mentioned, the extra divisor 1/2 in the case of an interaction is a convention.

EXAMPLE 12.3.2—From a randomized blocks experiment on sugar beets in Iowa the numbers of surviving plants per plot were counted as follows

Treatments	Blocks				Totals
	1	2	3	4	
None	183	176	291	254	904
Superphosphate, P	356	300	301	271	1228
Potash, K	224	258	244	217	943
P + K	329	283	308	326	1246
Totals	1092	1017	1144	1068	4321

(i) Compute the sums of squares for Blocks, Treatments, and Error. Verify that the Treatments $S.S.$ is 24,801, and the mean square for error is 1494

(ii) Compute the $S.S.$ for P, K, and the PK interaction. Verify that these add to the Treatments $S.S.$ and that the only significant effect is an increase of about 34% in plant number due to P. This result is a surprise, since P does not usually have marked effects on the number of sugar-beet plants.

(iii) Compute the factorial effect means from the individual treatment means with their $s.e. \sqrt{s^2/r}$, and verify that t -tests of the factorial effect means are identical to the F -tests in the analysis of variance.

EXAMPLE 12.3.3—We have seen how to calculate the factorial effect means (A), (B), and (AB) from the means (ab), (a), (b), and (1) of the individual treatment combinations. The process can be reversed: given the factorial effect means and the mean yield M of the experiment, we can recapture the means of the individual treatment combinations. Show that the equations are

$$\begin{aligned}
 (ab) &= M + \frac{1}{2} \{(A) + (B) + (AB)\} \\
 (a) &= M + \frac{1}{2} \{(A) - (B) - (AB)\} \\
 (b) &= M + \frac{1}{2} \{-(A) + (B) - (AB)\} \\
 (1) &= M + \frac{1}{2} \{-(A) - (B) + (AB)\}
 \end{aligned}$$

12.4—The 2^2 factorial when interaction is present. When interaction is present, the results of a 2^2 experiment require more detailed study. If both main effects are large, an interaction that is significant but much smaller than the main effects may imply merely that there is a minor variation in the effect of A according as B is at its higher or lower level, and vice versa. In this event, reporting of the main effects may still be an adequate summary. But in most cases we must revert to a report based on the 2×2 table.

Table 12.4.1 contains the results (slightly modified) of a 2^2 experiment in a completely randomized design. The factors were vitamin B_{12} (0, 5 mg.) and Antibiotics (0, 40 mg.) fed to swine. A glance at the totals for the four treatment combinations suggests that with no antibiotics, B_{12} had little or no effect (3.66 versus 3.57), apparently because intestinal flora utilized the B_{12} . With antibiotics present to control the flora, the effect of the vitamin was marked (4.63 versus 3.10). Looking at the table the other way, the antibiotics alone decreased gain (3.10 versus 3.57), perhaps by suppressing intestinal flora that synthesize B_{12} ; but with B_{12} added, the antibiotics produced a gain by decreasing the activities of unfavorable flora.

TABLE 12.4.1
 FACTORIAL EXPERIMENT WITH VITAMIN B₁₂ AND ANTIBIOTICS.
 AVERAGE DAILY GAIN OF SWINE (POUNDS)

Antibiotics	0		40 mg.				
	0	5 mg.	0	5 mg.			
B ₁₂							
	1.30	1.26	1.05	1.52			
	1.19	1.21	1.00	1.56			
	1.08	1.19	1.05	1.55			
Totals	3.57 (1)	3.66 <i>a</i>	3.10 <i>b</i>	4.63 <i>ab</i>	Factorial Effect Total	Factorial Effect Mean	<i>S E</i>
B ₁₂	-1	1	-1	1	1.62	0.270**	
Antibiotics	-1	-1	1	1	0.50	0.083*	±0.035
Interaction	1	-1	-1	1	1.44	0.240**	
Source of Variation		Degrees of Freedom		Sum of Squares		Mean Square	
Treatments		3		0.4124			
Error		8		0.0293		0.00366	

The summary of the results of this experiment is therefore presented in the form of a table of the means of the four treatment combinations, as shown below:

Antibiotics	0		40 mg.	
	0	5 mg.	0	5 mg.
B ₁₂				
Means	1.19	1.22	1.03	1.54

In the analysis of variance, s^2 is 0.00366, with 8 *d.f.* The *s.e.* of the difference between any two treatment means is $\sqrt{(2s^2/3)} = \pm 0.049$. You may verify that the decrease due to antibiotics when B₁₂ is absent, and the increases to each additive when the other is present, are all clearly significant.

If, instead, we begin by calculating the factorial effects, as shown in table 12.4.1, we learn from the factorial effect means that there is a significant interaction at the 1% level (0.240 ± 0.035). This immediately directs attention back to the four individual treatment totals or means, in order to study the nature of the interaction and seek an explanation. The main effects both happen to be significant, but are of no interest.

One way of describing the no-interaction situation is to say that the effects of the two factors are *additive*. To illustrate, suppose that the population mean for the (1) combination (neither factor present) is μ . Factor *A*, when present alone, changes the mean to $(\mu + \alpha)$: Factor *B*.

when present alone, to $(\mu + \beta)$. If both factors are present, and if their effects are additive, the mean will become $\mu + \alpha + \beta$.

With this model, the interaction effect is

$$(AB) = \frac{1}{2} [(ab) + (1) - (a) - (b)] = \frac{1}{2} [\mu + \alpha + \beta + \mu - \mu - \alpha - \mu - \beta] = 0$$

Presence of an interaction denotes that the effects are not additive.

With quantitative factors, this concept leads to two other possible explanations of an interaction found in an experiment. Sometimes their effects are additive, but on a transformed scale. The simplest example is that of multiplicative effects, in which a log transformation of the data before analysis (section 11.17) removes the interaction.

Secondly, if X_1 , X_2 represent the amounts of two factors in a treatment combination, it is natural to summarize the results by means of a *response function* or *response surface*, which predicts how the response Y varies as X_1 and X_2 are changed. If the effects are additive, the response function has the simple form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

A significant interaction is a warning that this model is not an adequate fit. The interaction effect may be shown to represent a term of the form $\beta_{12}X_1X_2$ in the response function. The presence of a term in X_1X_2 in the response function suggests that terms in X_1^2 and X_2^2 may also be needed to represent the function adequately. In other words, the investigator may require a quadratic response function. Since at least three levels of each variable are required to fit a quadratic surface, he may have to plan a larger factorial experiment.

EXAMPLE 12 4 1—Our use of the riboflavin data in section 12 3 as an example with no interaction might be criticized on two grounds: (1) a P value of 0.25 in the test for interaction in a small experiment suggests the possibility of an interaction that a larger experiment might reveal, (2) perhaps the effects are multiplicative in these data. If you analyze the logs of the data in table 12 3 1, you will find that the F -value for interaction is now only 0.7. Thus the assumption of zero interaction seems better grounded on a log scale than on the original scale.

12.5—The general two-factor experiment. Leaving the special case of two levels per factor, we now consider the general arrangement with a levels of the first factor and b levels of the second. As before, the layout of the experiment may be completely randomized, randomized blocks, or any other standard plan.

With a levels, the main effects of A in the analysis of variance now have $(a - 1)$ d.f., while those of B have $(b - 1)$ d.f. Since there are ab treatment combinations, the Treatments S.S. has $(ab - 1)$ d.f. Consequently, there remain

$$(ab - 1) - (a - 1) - (b - 1) = ab - a - b + 1 = (a - 1)(b - 1)$$

d.f., which may be shown to represent the AB interactions. In the 2×2

factorial, in which the AB interaction had only one $d.f.$, the comparison corresponding to this $d.f.$ was called *the* AB interaction. In the general case, the AB interaction represents a set of $(a - 1)(b - 1)$ independent comparisons. These can be subdivided into single comparisons in many ways.

In deciding how to subdivide the AB sum of squares, the investigator is guided by the questions that he had in mind when planning the experiment. Any comparison among the levels of A is estimated independently at each of the b levels of B . For a comparison that is of particular interest, the investigator may wish to examine whether the level of B affects these estimates. The sum of squares of deviations of the estimates, with the appropriate divisor, is a component of the AB interaction, with $(b - 1)$ $d.f.$, which may be isolated and tested against the Error mean square. Incidentally, since the main effect of A represents $(a - 1)$ independent comparisons, these components of the AB interaction jointly account for $(a - 1)(b - 1)$ $d.f.$ and will be found to sum to the sum of squares for AB .

As an illustration, the data in table 12.5.1 show the gains in weight of male rats under six feeding treatments in a completely randomized experiment. The factors were:

A (3 levels): Source of protein: Beef, Cereal, Pork

B (2 levels): Level of protein: High, Low

Often the investigator has decided in advance how to subdivide the comparisons that represent main effects and interactions. In more ex-

TABLE 12 5 1
GAINS IN WEIGHT (GRAMS) OF RATS UNDER SIX DIETS

	High Protein			Low Protein		
	Beef	Cereal	Pork	Beef	Cereal	Pork
	73	98	94	90	107	49
	102	74	79	76	95	82
	118	56	96	90	97	73
	104	111	98	64	80	86
	81	95	102	86	98	81
	107	88	102	51	74	97
	100	82	108	72	74	106
	87	77	91	90	67	70
	117	86	120	95	89	61
	111	92	105	78	58	82
Totals	1,000	859	995	792	839	787
Source of Variation	Degrees of Freedom		Sum of Squares	Mean Square		F
Treatments	5		4,613 0			
A (Source of protein)	2		266 5	133 2		0 6
B (Level of protein)	1		3,168 3	3,168 3		14 8**
AB	2		1,178 2	589 1		2 7
Error	54		11,585 7	214 6		

ploratory situations, it is customary to start with a breakdown of the Treatments *S.S.* into the *S.S.* for *A*, *B*, and *AB*. This has been done in table 12.5.1. Looking at the main effects of *A*, the three sources of protein show no differences in average rates of gain ($F = 0.6$), but there is a clear effect of level of protein ($F = 14.8$), the gain being about 18% larger with the High level.

For *AB*, the value of F is 2.7, between the 10% and the 5% level. In the general two-factor experiment and in more complex factorials, it often happens that a few of the comparisons comprising the main effects have substantial interactions while the majority of the comparisons have negligible interactions. Consequently, the F -test of the *AB* interaction sum of squares as a whole is not a good guide as to whether interactions can be ignored. It is well to look over the two-way table of treatment totals or means before concluding that there are no interactions, particularly if F is larger than 1.

Another working rule tested by experience in a number of areas is that large main effects are more likely to have interactions than small ones. Consequently, we look particularly at the effects of *B*, Level of protein. From the treatment totals in table 12.5.1 we see that high protein gives large gains over low protein for beef and pork, but only a small gain for cereal. This suggests a breakdown into: (1) Cereal versus the average of Beef and Pork, and (2) Beef versus Pork. This subdivision is a natural one, since Beef and Pork are animal sources of protein while Cereal is a vegetable source, and would probably be planned from the beginning in this type of experiment.

Table 12.5.2 shows how this breakdown is made by means of five single comparisons. Study the coefficients for each comparison carefully, and verify that the comparisons are mutually orthogonal. In the lower part of the table the divisors required to convert the squares of the factorial effect totals into sums of squares in the analysis of variance are given. Each divisor is n times the sum of squares of the coefficients in the comparison ($n = 10$). As anticipated, the interaction of the animal versus vegetable comparison with level of protein is significant at the 5% level. There is no sign of a difference between Beef and Pork at either level.

The principal results can therefore be summarized in the following 2×2 table of means.

Mean Rat Gains in Weight per Week (Grams)				
Level of Protein	Source of Protein		Difference	<i>S E</i>
	Animal	Vegetable		
High	99.8	85.9	+13.9*	±5.67
Low	79.0	83.9	- 4.9	±5.67
Difference	+20.8**	+ 2.0		
<i>S E</i>	± 4.6	± 6.5		

TABLE 12.5.2
SUBDIVISION OF THE S.S. FOR MAIN EFFECTS AND INTERACTIONS

Comparisons (Treatment Totals)	High Protein			Low Protein			Factorial Effect Total
	Beef 1000	Cereal 859	Pork 995	Beef 792	Cereal 839	Pork 787	
Level of protein	+1	+1	+1	-1	-1	-1	436
Animal vs. vegetable	+1	-2	+1	+1	-2	+1	178
Interaction with level	+1	-2	+1	-1	+2	-1	376
Beef vs. pork	+1	0	-1	+1	0	-1	10
Interaction with level	+1	0	-1	-1	0	+1	0

Comparison	Divisor for S.S.	Degrees of Freedom	Sum of Squares	Mean Square
Level of protein	60	1	3168.3**	
Animal vs. vegetable	120	1	264.0	
Interaction with level	120	1	1178.1*	
Beef vs. pork	40	1	2.5	
Interaction with level	40	1	0.0	
Error		54		214.6

As a consequence of the interaction, the animal proteins gave substantially greater gains in weight than cereal protein at the high level, but showed no superiority to cereal protein at the low level.

12.6—Response curves. Frequently, the levels of a factor represent increasing amounts X of some substance. It may then be of interest to examine whether the response Y to the factor has a linear relation to the amount X . An example has already been given in section 11.8, p. 313, in which the linear regression of yield of millet on width of spacing of the rows was worked out for a Latin square experiment. If the relation between Y and X is curved, a more complex mathematical expression is required to describe it. Sometimes the form of this expression is suggested by subject-matter knowledge. Failing this, a polynomial in X is often used as a descriptive equation.

With equally spaced levels of X , auxiliary tables are available that facilitate the fitting of these polynomials. The tables are explained fully in section 15.6 (p. 460). An introduction is given here to enable them to be used in the analysis of factorial experiments. The tables are based essentially on an ingenious coding of the values of X , X^2 , and so on.

With three levels, the values of X are coded as $-1, 0, +1$, so that they sum to 0. If Y_1, Y_2, Y_3 are the corresponding response totals over n replicates, the linear regression coefficient b_1 is $\Sigma XY/n\Sigma X^2$, or $(Y_3 - Y_1)/2n$. The values of X^2 are 1, 0, 1. Subtracting their mean $2/3$ so that they add to 0 gives $1/3, -2/3, 1/3$. Multiplying by 3 in order to have whole numbers, we get the coefficients 1, $-2, 1$. In its coded form, this variable is $X_2 = 3X^2 - 2$. The regression coefficient of Y on X_2 is

$b_2 = \Sigma X_2 Y / n \Sigma X_2^2$, or $(Y_3 - 2Y_2 + Y_1)/6n$. The equation for the parabola fitted to the level means of Y is

$$\hat{Y} = \bar{Y} + b_1 X + b_2 X_2 \quad (12.6.1)$$

With four levels of X , they are coded $-3, -1, +1, +3$, so that they are whole numbers adding to 0. The values of X^2 are 9, 1, 1, 9, with mean 5. Subtracting the mean gives $+4, -4, -4, +4$, which we divide by 4 to give the coefficients $+1, -1, -1, +1$ for the parabolic component. These components represent the variable $X_2 = (X^2 - 5)/4$. The fitted parabola has the same form as (12.6.1), where

$$b_1 = (3Y_4 + Y_3 - Y_2 - 3Y_1)/20n : b_2 = (Y_4 - Y_3 - Y_2 + Y_1)/4n,$$

the Y_i being level totals. For the cubic component (term involving X^3) a more elaborate coding is required to make this orthogonal to X and X_2 . The resulting coefficients are $-1, +3, -3, +1$.

By means of these polynomial components, the *S.S.* for the main effects of the factor can be subdivided into linear, quadratic, cubic components, and so on. Each *S.S.* can be tested against the Error mean square as a guide to the type of polynomial that describes the response curve. By rule 11.6.1, the contribution of any component $\Sigma \lambda_i Y_i$ to the *S.S.* is $(\Sigma \lambda_i Y_i)^2 / n \Sigma \lambda_i^2$. If the component is computed from the level means, as in the following illustration, the divisor is $(\Sigma \lambda^2)/n$.

Table 12.6.1 presents the mean yields of sugar (cwt. per acre) in an experiment (4) on beet sugar in which a mixture of fertilizers was applied at four levels (0, 4, 8, 12 cwt. per acre).

TABLE 12.6.1
LINEAR, QUADRATIC, AND CUBIC COMPONENTS OF RESPONSE CURVE

Mean Yields	Mixed Fertilizers (Cwt. Per Acre)				Component	Sum of Squares	<i>F</i>
	0	4	8	12			
	34.8	41.1	42.6	41.8			
Linear	-3	-1	+1	+3	+22.5	202.5	17.0**
Quadratic	+1	-1	-1	+1	-7.1	100.8	8.5*
Cubic	-1	+3	-3	+1	+2.5	2.5	0.2
Total = Sum of Squares for Fertilizers =						305.8	

Error mean square (16 *d.f.*) = 11.9

Since each mean was taken over $n = 8$ replicates, the divisors are $20/8 = 2.5$ for the linear and cubic components and $4/8 = 0.5$ for the quadratic component. The Error mean square was 11.9 with 16 *d.f.* The positive linear component and the negative quadratic component are both significant, but the cubic term gives an *F* less than 1. The conclusions are: (i) mixed fertilizers produced an increase in the yield of sugar, (ii) the rate of increase fell off with the higher levels.

To fit the parabola, we compute from table 12.6.1,

$$\bar{Y} = 40.08 : b_1 = +22.5/20 = 1.125 : b_2 = -7.1/4 = -1.775$$

The fitted parabola is therefore

$$\hat{Y} = 40.08 + 1.125X - 1.775X_2, \quad (12.6.2)$$

where \hat{Y} is an estimated mean yield. The estimated yields for 0, 4, 8, 12 cwt. of fertilizers are 34.93, 40.73, 42.98, 41.68 cwt. per acre. Like the observed means, the parabola suggests that the dressing for maximum yield is around 8 cwt. per acre.

Table 12.6.2 shows the coefficients for the polynomial components and the values of $\Sigma\lambda^2$ for factors having from 2 to 7 levels. With k levels a polynomial of degree $(k - 1)$ can be made to fit the k responses exactly.

TABLE 12.6.2
COEFFICIENTS AND DIVISORS FOR SETS OF ORTHOGONAL COMPONENTS IN REGRESSION
IF X IS SPACED AT EQUAL INTERVALS

Degree of Polynomial	Comparison	Number of Levels							Divisor $\Sigma\lambda^2$
		1	2	3	4	5	6	7	
1	Linear	-1	+1						2
2	Linear Quadratic	-1	0	+1					2 6
3	Linear Quadratic Cubic	-3	-1	+1	+3				20 4 20
4	Linear Quadratic Cubic Quartic	-2	-1	0	+1	+2			10 14 10 70
5	Linear Quadratic Cubic Quartic Quintic	-5	-3	-1	+1	+3	+5		70 84 180 28 252
6	Linear Quadratic Cubic Quartic Quintic Sextic	-3	-2	-1	0	+1	+2	+3	28 84 6 154 84 924

EXAMPLE 12.6.1—In the same sugar-beet experiment, the mean yield of tops (green matter) for 0, 4, 8, 12 cwt. fertilizers were 9.86, 11.58, 13.95, 14.95 cwt. per acre. The Error mean square was 0.909. Show that: (i) only the linear component is significant, there being no apparent decline in response to the higher applications, (ii) the S.S. for the linear, quadratic, and cubic components sum to the S.S. between levels, 127.14 with 3 d.f. Remember that the means are over 8 replicates.

352 Chapter 12: Factorial Experiments

EXAMPLE 12 6 2—From the results for the parabolic regression on yield of sugar the estimated optimum dressing can be computed by calculus. From equation 12 6 2 the fitted parabola is

$$\hat{Y} = 40.08 + 1.125X_1 - 1.775X_2,$$

where $X_2 = (X^2 - 5)/4$. Thus

$$\hat{Y} = 40.08 + 1.125X - 0.444(X^2 - 5)$$

Differentiating, we find a turning value at $X = 1.125/0.888 = 1.27$ on the coded scale. You may verify that the estimated maximum sugar yield is 43.0 cwt., for a dressing of 8.5 cwt fertilizer.

12.7—Response curves in two-factor experiments. Either or both factors may be quantitative and may call for the fitting of a regression as described in the previous section. As an example with one quantitative

TABLE 12 7.1
YIELD OF COWPEA HAY (POUNDS PER 1/100 MORGEN PLOT) FROM THREE VARIETIES

Varieties	Spacing (In.)	Blocks				Sum
		1	2	3	4	
I	4	56	45	43	46	190
	8	60	50	45	48	203
	12	66	57	50	50	223
II	4	65	61	60	63	249
	8	60	58	56	60	234
	12	53	53	48	55	209
III	4	60	61	50	53	224
	8	62	68	67	60	257
	12	73	77	77	65	292
Sum		555	530	496	500	2,081

Varieties	Spacings			
	4	8	12	
I	190	203	223	616
II	249	234	209	692
III	224	257	292	773
Sum	663	694	724	2,081

	Degrees of Freedom	Sum of Squares	Mean Square
Blocks	3	255.64	
Varieties, V	2	1027.39	513.70**
Spacings, S	2	155.06	77.53*
Interactions, V/S	4	765.44	191.36**
Error	24	424.11	17.67

factor, table 12.7.1 shows the yields in a 3×3 factorial on hay (5), one factor being three widths of spacing of the rows, the other being three varieties.

The original analysis of variance, at the foot of table 12.7.1, reveals marked VS (variety \times spacing) interactions. The table of treatment combination totals immediately above shows that there is an upward trend in yield with wider spacing for varieties I and III but an opposite trend with variety II. This presumably accounts for the large VS mean square and warns that no useful overall statements can be made from the main effects.

To examine the trends of yield Y on spacing X , the linear and quadratic components are calculated for each variety, table 12.7.2. The factorial effect totals for these components are computed first, then the corresponding sums of squares. Note the following results from table 12.7.2:

(i) As anticipated, the linear slopes are positive for varieties I and III and negative for variety II.

(ii) The linear trend for each variety is significant at the 1% level, while no variety shows any sign of curvature, when tested against the Error mean square of 17.67.

TABLE 12.7.2
LINEAR AND QUADRATIC COMPONENTS FOR EACH VARIETY IN COWPEA EXPERIMENT

Linear Quadratic	4"	8"	12"	Totals for Components	
	-1 +1	0 -2	+1 +1	Linear	Quadratic
Variety I	190	203	223	33	7
Variety II	249	234	209	-40	-10
Variety III	224	257	292	68	2
Sum	663	694	724	61	-1

Contributions to Sums of Squares

Variety I	Linear, $\frac{(33)^2}{(4)(2)} = 136.12^{**}$	Quadratic, $\frac{(7)^2}{(4)(6)} = 2.04$
II	$\frac{(-40)^2}{(4)(2)} = 200.00^{**}$	$\frac{(-10)^2}{(4)(6)} = 4.17$
III	$\frac{(68)^2}{(4)(2)} = 578.00^{**}$	$\frac{(2)^2}{(4)(6)} = 0.17$
Total	914.12	6.38

Verification $914.12 + 6.38 = 155.06 + 765.44 (=S + SV), f = 6$

(iii) The sum of these six *S.S.* is identical with the *S.S.* for spacings and interactions combined, 920.50.

(iv) If the upward trends for varieties I and III are compared, the trend for variety III will be found significantly greater.

To summarize, the varieties have linear trends on spacing which are not the same. Apparently I and III have heavy vegetative growth which requires more than 12" spacing for maximum yield. In a further experiment the spacings tested for varieties I and III should differ from those for II.

EXAMPLE 12.7.1—In the variety \times spacing experiment, verify the statement that the linear regression of yield on width of spacing is significantly greater for variety III than for variety I.

EXAMPLE 12.7.2—If the primary interest in this experiment were in comparing the varieties when each has its highest-yielding spacing, we might compare the totals 223 (I), 249 (II), and 292 (III). Show that the optimum for III exceeds the others at the 1% level

12.8—Example of a response surface. We turn now to a 3×4 experiment in which there is regression in each factor. The data are from the Foods and Nutrition Section of the Iowa Agricultural Experiment Station (6). The object was to learn about losses of ascorbic acid in snapbeans stored at 3 temperatures for 4 periods, each 2 weeks longer than the preceding. The beans were all harvested under uniform conditions before eight o'clock one morning. They were prepared and quick-frozen before noon of the same day. Three packages were assigned at random to each of the 12 treatments and all packages were stored at random positions in the locker, a completely randomized design.

The sums of 3 ascorbic acid determinations are recorded in table 12.8.1. It is clear that the concentration of ascorbic acid decreases with

TABLE 12.8.1
SUM OF THREE ASCORBIC ACID DETERMINATIONS (MG/100 G.) FOR EACH OF 12 TREATMENTS
IN A 3×4 FACTORIAL EXPERIMENT ON SNAPBEANS

Temperature, F.°	Weeks of Storage				Sum
	2	4	6	8	
0	45	47	46	46	184
10	45	43	41	37	166
20	34	28	21	16	99
Sum	124	118	108	99	449

	Degrees of Freedom	Sum of Squares	Mean Square
Temperature, <i>T</i>	2	334.39	
Two-week Period, <i>P</i>	3	40.53	
Interaction, <i>TP</i>	6	34.05	
Error*	24		0.706

* Error (packages of same treatment) was calculated from original data not recorded here

higher storage temperatures and, except at 0°, with storage time. It looks as if the rate of decrease with temperature is not linear and not the same for the several storage periods. These conclusions, suggested by inspection of table 12.8.1, will be tested in the following analysis:

One can look first at either temperature or period; we chose temperature. At each period the linear and quadratic temperature comparisons (-1, 0, +1; +1, -2, +1) are calculated:

Weeks of Storage	2	4	6	8	Total
Linear, T_L	-11	-19	-25	-30	-85
Quadratic, T_Q	-11	-11	-15	-12	-49

The downward slopes of the linear regressions get steeper with time. This will be examined later. At present, calculate sums of squares as follows:

$$T_L = \frac{(-85)^2}{(12)(2)} = 301.04^{**}$$

$$T_Q = \frac{(-49)^2}{(12)(6)} = 33.35^{**}$$

The sum is the sum of squares for T , $301.04 + 33.35 = 334.39$. Significance in each effect is tested by comparison with the Error mean square, 0.706. Evidently the regressions are curved, the parabolic comparison being significant; quality decreases with accelerated rapidity as the temperature increases. (Note the number of replications in each temperature total, 4 periods times 3 packages = 12.)

Are the regressions the same for all periods? To answer this, calculate the interactions of the linear and the quadratic comparisons with period. The sums of squares for these interactions are:

$$T_LP : \frac{(-11)^2 + \dots + (-30)^2}{(3)(2)} - T_L = 33.46^{**} \quad (3 \text{ d.f.})$$

$$T_QP : \frac{(-11)^2 + \dots + (-12)^2}{(3)(6)} - T_Q = 0.59 \quad (3 \text{ d.f.})$$

Rule 12.8.1. These calculations follow from a new rule. If a comparison L_i has been computed for k different levels of a second factor, the Interaction S.S. of this comparison with the second factor is

$$\frac{\sum L_i^2}{n(\sum \lambda^2)} - \frac{(\sum L_i)^2}{kn(\sum \lambda^2)} \quad (i = 1, 2, \dots, k)$$

with $(k - 1)$ d.f. Further, the term $(\sum L_i)^2 / kn(\sum \lambda^2)$ is the overall S.S. (1 d.f.) for this comparison. The sum of T_LP and T_QP is equal to the sum of squares for TP . The linear regressions decrease significantly with

period (length of storage) but the quadratic terms may be the same for all periods, since the mean square for T_QP , $0.59/3 = 0.20$, is smaller than the Error mean square.

Turning to the sums for the 4 periods, calculate the 3 comparisons:

Sums	124	118	108	99	Comparison	Sum of Squares
Linear, P_L	-3	-1	+1	+3	-85	40.14**
Quadratic, P_Q	+1	-1	-1	+1	-3	0.25
Cubic, P_C	-1	+3	-3	+1	5	0.14
Sum = Sum of Squares for Periods						40.53

This indicates that the population regressions on period may be linear, the mean squares 0.25 for P_Q and 0.14 for P_C being both less than 0.706, the Error mean square.

We come now to the new feature of this section, the regressions of T_L and T_Q on period. T_L , the downward slope of the vitamin with temperature, has been calculated for each period; the question is, in what manner does T_L change with period?

For this question, we can work out the linear, quadratic, and cubic components of the regression of T_L on period, just as was done above for the sums over the 4 periods.

T_L	-11	-19	-25	-30	Comparison	Divisor	Sum of Squares
Linear, $T_L P_L$	-3	-1	+1	+3	-63	(3)(2)(20)	33.08**
Quadratic, $T_L P_Q$	+1	-1	-1	+1	3	(3)(2)(4)	0.38
Cubic, $T_L P_C$	-1	+3	-3	+1	-1	(3)(2)(20)	0.01
Sum = Sum of Squares for $T_L P$							33.47

Rule 12.8.2. Note the rule for finding the divisors. For each individual T_L (-11, -19, etc.) the divisor was (2)(3). We now have a comparison among these T_L 's, bringing in a further factor 20 = $3^2 + 1^2 + 1^2 + 3^2$ in $T_L P_L$. Thus the S.S. $33.08 = (-63)^2/120$. The sum of the three regression sums of squares is 33.47, which equals $T_L P$. From the tests of the linear, quadratic, and cubic components, we conclude that the linear regression on temperature decreases linearly with length of storage.

Proceeding in the same way with T_Q :

$$T_Q P_L = \frac{(-7)^2}{(3)(6)(20)} = 0.14$$

$$T_Q P_Q = \frac{(3)^2}{(3)(6)(4)} = 0.12$$

$$T_Q P_C = \frac{(11)^2}{(3)(6)(20)} = 0.34$$

TABLE 12.8.2
ANALYSIS OF VARIANCE OF ASCORBIC ACID IN SNAP BEANS

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Temperature:	(2)	(334.39)	
T_L	1		301.04**
T_Q	1		33.35**
Period:	(3)	(40.53)	
P_L	1		40.14**
P_Q	1		0.25
P_C	1		0.14
Interaction:	(6)	(34.05)	
$T_L P_L$	1		33.08**
$T_L P_Q$	1		0.38
$T_L P_C$	1		0.01
$T_Q P_L$	1		0.14
$T_Q P_Q$	1		0.12
$T_Q P_C$	1		0.34
Error	24		0.706

The sum is $T_Q P = 0.60$. Clearly there is no change in T_Q with period. The results are collected in table 12.8.2

In summary, T_L and T_Q show that the relation of ascorbic acid to temperature is parabolic, the rate of decline increasing as storage time lengthens ($T_L P_L$). The regression on period is linear, sloping downward more rapidly as temperature increases. In fact, you will note in table 12.8.1 that at the coldest temperature, 0°F, there is no decline in amount of ascorbic acid with additional weeks of storage.

These results can be expressed as a mathematical relation between ascorbic acid Y , storage temperature T , and weeks of storage W . As we have seen, we require terms in T_L , T_Q , P_L , and $T_L P_L$ in order to describe the relation adequately. It is helpful to write down these polynomial coefficients for each of the 12 treatment combinations, as shown in table 12.8.3.

For the moment, think of the mathematical relation as having the form

$$\hat{Y} = \bar{Y} + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$$

where \hat{Y} is the predicted ascorbic acid *total* over 3 replications, while $X_1 = T_L$, $X_2 = T_Q$, $X_3 = P_L$, and $X_4 = T_L P_L$. The regression coefficient $b_i = \Sigma X_i Y / \Sigma X_i^2$. The quantities $\Sigma X_i Y$, which were all obtained in the earlier analysis, are given at the foot of table 12.8.3, as well as the divisors ΣX_i^2 . Hence, the relation is as follows:

$$\hat{Y} = 37.417 - 10.625X_1 - 2.042X_2 - 1.417X_3 - 1.575X_4 \quad (12.8.1)$$

Since the values of the X_i are given in table 12.8.3, the predicted values \hat{Y} are easily computed for each treatment combination. For example, for 0°F. and 2 weeks storage.

TABLE 12.8.3
CALCULATION OF THE RESPONSE SURFACE

Temp.	Weeks	Y Totals	$T_L =$ $0.1(T-10)$	$T_Q =$ $3T_L^2 - 2$	$P_L =$ $W-5$	$T_L P_L =$ $0.1(T-10)(W-5)$	\hat{Y}
0°	2	45	-1	+1	-3	+3	45.53
	4	47	-1	+1	-1	+1	45.84
	6	46	-1	+1	+1	-1	46.16
	8	46	-1	+1	+3	-3	46.47
10°	2	45	0	-2	-3	0	45.75
	4	43	0	-2	-1	0	42.92
	6	41	0	-2	+1	0	40.08
	8	37	0	-2	+3	0	37.25
20°	2	34	+1	+1	-3	-3	33.73
	4	28	+1	+1	-1	-1	27.74
	6	21	+1	+1	+1	+1	21.76
	8	16	+1	+1	+3	+3	15.77
$\Sigma X_i Y$		449	-85	-49	-85	-63	
Divisor for b		12	8	24	60	40	

$$\hat{Y} = 37.417 - (10.625)(-1) - 2.042(+1) - 1.417(-3) - (1.575)(+3) = 45.53.$$

as shown in the right-hand column of table 12.8.3.

By decoding, we can express the prediction equation (12.8.1) in terms of T (°F.) and W (weeks). You may verify that the relations between $X_1(T_L)$, $X_2(T_Q)$, $X_3(P_L)$, $X_4(T_L P_L)$ and T and W are as given at the top of table 12.8.3. After making these substitutions and dividing by 3 so that the prediction refers to the ascorbic acid *mean* per treatment combination, we have

$$\hat{Y} = 15.070 + 0.3167T - 0.92042T^2 + 0.0528W - 0.05250TW \quad (12.8.2)$$

Geometrically, a relation of this type is called a *response surface*, since we have now a relation in three dimensions Y , T , and W . With quantitative factors, the summarization of the results by a response surface has proved highly useful, particularly in industrial research. If the objective of the research is to maximize Y , the equation shows the combinations of levels of the factors that give responses close to the maximum. Further accounts of this technique, with experimental plans specifically constructed for fitting response surfaces, are given in (7) and (8). The analysis in this example is based on (6).

A word of warning. In the example we fitted a multiple regression of Y on four variables X_1 , X_2 , X_3 , X_4 . The methods by which the regression coefficients b_i were computed apply only if the X_i are mutually orthogonal, as was the case here. General methods are presented in chapter 13.

12.9—Three-factor experiments; the 2^3 . The experimenter often requires evidence about the effects of 3 or more factors in a common environment. The simplest arrangement is that of 3 factors each at 2 levels, the $2 \times 2 \times 2$ or 2^3 experiment. The eight treatment combinations may be tried in any of the common experimental designs.

The data in table 12.9.1 are extracted from an unpublished randomized blocks experiment (9) to learn the effect of two supplements to a corn ration for feeding pigs. The factors were as follows:

Lysine (*L*) : 0 and 0.6%.

Soybean meal (*P*) : Amounts added to supply 12% and 14% protein.

Sex (*S*) : Male and Female.

TABLE 12.9.1
AVERAGE DAILY GAINS OF PIGS IN 2^3 FACTORIAL ARRANGEMENT OF TREATMENTS.
RANDOMIZED BLOCKS EXPERIMENT

Lysine %	Protein %	Sex	Replications (Blocks)								Treatment Sum	Sum for 2 Sexes
			1	2	3	4	5	6	7	8		
0	12	M	1.11	0.97	1.09	0.99	0.85	1.21	1.29	0.96	8.47	17.08
		F	1.03	0.97	0.99	0.99	0.99	1.21	1.19	1.24	8.61	
	14	M	1.52	1.45	1.27	1.22	1.67	1.24	1.34	1.32	11.03	21.74
		F	1.48	1.22	1.53	1.19	1.16	1.57	1.13	1.43	10.71	
	12	M	1.22	1.13	1.34	1.41	1.34	1.19	1.25	1.32	10.20	19.34
		F	0.87	1.00	1.16	1.29	1.00	1.14	1.36	1.32	9.14	
	14	M	1.38	1.08	1.40	1.21	1.46	1.39	1.17	1.21	10.30	19.93
		F	1.09	1.09	1.47	1.43	1.24	1.17	1.01	1.13	9.63	
Replication Sum			9.70	8.91	10.25	9.73	9.71	10.12	9.74	9.93		78.09
			Degrees of Freedom				Sum of Squares			Mean Square		
Replications			7				0.1411					
Treatments			7				0.7986			0.1141**		
Error			49				1.0994			0.0224		

With three factors there are three main effects, *L*, *P*, and *S*; three two-factor interactions, *SP*, *SL*, and *LP*; and a *three-factor interaction SLP*. The comparisons representing the factorial effect totals are set out in table 12.9.2. The coefficients for the main effects and the two-factor interactions should present no difficulty, these being the same as in a 2^2 factorial. A useful rule in the 2^n series is that the coefficients for any two-factor interaction like *SP* are the products of the corresponding coefficients for the main effects *S* and *P*.

The new term is the three-factor interaction *SLP*. From table 12.9.2 the *SP* interaction (apart from its divisor) can be estimated at the higher level of *L* as

$$10.20 - 9.14 - 10.30 + 9.63 = +0.39$$

TABLE 12.9.2
SEVEN COMPARISONS IN 2^3 FACTORIAL EXPERIMENT ON PIGS

Effects	Lysine = 0				Lysine = 0.6%				Factorial Effect Total	Sum of Squares
	<i>P</i> = 12%		<i>P</i> = 14%		<i>P</i> = 12%		<i>P</i> = 14%			
	M	F	M	F	M	F	M	F		
	8.47	8.61	11.03	10.71	10.20	9.14	10.30	9.63		
Sex, <i>S</i>	−1	+1	−1	+1	−1	+1	−1	+1	−1.91	0.0570
Protein, <i>P</i>	−1	−1	+1	+1	−1	−1	+1	+1	5.25	0.4307**
<i>SP</i>	+1	−1	−1	+1	+1	−1	−1	+1	−0.07	0.0001
Lysine, <i>L</i>	−1	−1	−1	−1	+1	+1	+1	+1	0.45	0.0032
<i>SL</i>	+1	−1	+1	−1	−1	+1	−1	+1	−1.55	0.0375
<i>PL</i>	+1	+1	−1	−1	−1	−1	+1	+1	−4.07	0.2588**
<i>SPL</i>	−1	+1	+1	−1	+1	−1	−1	+1	0.85	0.0113
Total										0.7986

An independent estimate at the lower level of L is

$$8.47 - 8.61 - 11.03 + 10.71 = -0.46$$

The sum of these two quantities, -0.07 , is the factorial effect total for SP . Their difference, $+0.39 - (-0.46) = +0.85$, measures the effect of the level of L on the SP interaction. If we compute in the same way the effect of P on the SL interaction, or of S on the PL interaction, the quantity $+0.85$ is again obtained. It is called the factorial effect total for SLP . Such interactions are rather difficult to grasp. Fortunately, they are often negligible except in experiments that have large main effects. A significant three-factor interaction is a sign that the corresponding 3-way table of means must be examined in the interpretation of the results.

As usual, the square of each factorial effect total is divided by $n(\Sigma\lambda^2)$, where $n = 8$ and $\Sigma\lambda^2 = 8$, the denominator being 64 in every case. As a check, the total of the sums of squares for the factorial effects in table 12.9.2 must add to the Treatments sum of squares in table 12.9.1, 0.7986.

The only significant effects are the main effect of P and the PL interaction. The totals for the $P \times L$ 2-way table are shown in the right hand column of table 12.9.1. With no added lysine, the higher level of protein gave a substantially greater daily gain than the lower level, but with added lysine, this gain was quite small. The result is not surprising, since soybean meal contains lysine. Lysine increased the rate of gain at the lower level of protein but decreased it at the higher level.

In view of these results there is no interest in the main effects of P or of L . The experimenter has learned that gains can be increased either by a heavier addition of soybean meal or by the addition of lysine, whichever is more profitable: he should not add both. The absence of any interactions involving S gives some assurance that these results hold for both males and females.

The 2nd factorial experiment has proved a potent research weapon in many fields. For further instruction on analysis, with examples, see (7) (8), and (10).

12.10—Three-factor experiments; a $2 \times 3 \times 4$. This section illustrates the general method of analysis for a three-factor experiment. The data come from the experiment drawn on in the previous section. The factors were Lysine (4 levels), Methionine (3 levels), and Soybean Meal (2 levels of protein), as food supplements to corn in pig feeding. Only the males in two replications are used. This makes a $2 \times 3 \times 4$ factorial arrangement of treatments in a randomized blocks design. Table 12.10.1 contains the data, with the computations for the analysis of variance given in detail.

1. First form the sums for each treatment and replication, and compute the total *S.S.* and the *S.S.* for treatments, replications, and error (by subtraction).

2. For each pair of factors, form a two-way table of sums. From the $L \times M$ table (table *A*), obtain the total *S.S.* (11 *df.*) and the *S.S.* for *L* and *M*. The *S.S.* for the *LM* interactions is found by subtraction. The $M \times P$ table supplies the *S.S.* for *M* (already obtained), for *P*, and for the *MP* interactions (by subtraction). The $L \times P$ table provides the *S.S.* for the *LP* interactions.

3. From the *S.S.* for treatments subtract the *S.S.* for *L*, *M*, *P*, *LM*, *MP*, and *LP* to obtain that for the *LMP* three-factor interactions.

The analysis of variance appears in table 12.10.2, and a further examination of the results in examples 12.10.1 to 12.10.3.

EXAMPLE 12.10.1—In table 12.10.2, for *L*, *M*, *MP*, and *LMP* the sums of squares are all so small that no single degree of freedom isolated from them could reach significance. But *LM* and *LP* deserve further study.

In the *LM* summary table *A* in table 12.10.1, there is some evidence of interaction though the overall test on 6 degrees of freedom doesn't detect it. Let us look at the line effects. First, calculate M_L (−1, 0, +1) for each level of lysine

$$-0.08, -0.27, 0.57, 1.07$$

Next, take the linear effect of lysine (−3, −1, +1, +3) in these M_L ; the result, 4.29. Final application of Rule 12.8.2 yields the sum of squares

$$L_L M_L = \frac{(4.29)^2}{(4)(2)(20)} = 0.1150,$$

which is just short of significance at the 5% level. None of the other 5 comparisons is significant. In the larger experiment of which this is a part, $L_L M_L$ was significant. What interpretation do you suggest?

EXAMPLE 12.10.2—In the *LP* summary table *C*, the differences between 14% and 12%

$$2.15, 2.07, 0.29, 0.56,$$

suggest an interaction: the beneficial effect of the higher level of protein decreases as more lysine is added. By applying the multipliers −3, −1, +1, +3, to the above figures, we obtain the $L_L P_L$ effect total = −6.55. By Rule 12.8.2,

TABLE 12.10.1
THREE-FACTOR EXPERIMENT ($2 \times 3 \times 4$) IN RANDOMIZED BLOCKS. AVERAGE DAILY
GAINS OF PIGS FED VARIOUS PERCENTAGES OF SUPPLEMENTARY LYSINE,
METHIONINE, AND PROTEIN

Lysine, <i>L</i>	Methionine, <i>M</i>	Protein, <i>P</i>	Replications (Blocks)		Treatment Total
			1	2	
0	0	12	1.11	0.97	2.08
		14	1.52	1.45	2.97
	0.025	12	1.09	0.99	2.08
		14	1.27	1.22	2.49
	0.050	12	0.85	1.21	2.06
		14	1.67	1.24	2.91
0.05	0	12	1.30	1.00	2.30
		14	1.55	1.53	3.08
	0.025	12	1.03	1.21	2.24
		14	1.24	1.34	2.58
	0.050	12	1.12	0.96	2.08
		14	1.76	1.27	3.03
0.10	0	12	1.22	1.13	2.35
		14	1.38	1.08	2.46
	0.025	12	1.34	1.41	2.75
		14	1.40	1.21	2.61
	0.050	12	1.34	1.19	2.53
		14	1.46	1.39	2.85
0.15	0	12	1.19	1.03	2.22
		14	0.80	1.29	2.09
	0.025	12	1.36	1.16	2.52
		14	1.42	1.39	2.81
	0.050	12	1.46	1.03	2.49
		14	1.62	1.27	2.89
Total			31.50	28.97	60.47

Computations:

1. $C = (60.47)^2/48 = 76.1796$
2. Total: $1.11^2 + 0.97^2 + \dots + 1.62^2 + 1.27^2 - C = 2.0409$
3. Treatments: $(2.08^2 + 2.97^2 + \dots + 2.89^2)/2 - C = 1.2756$
4. Replications: $(31.50^2 + 28.97^2)/24 - C = 0.1334$
5. Error: $2.0409 - (1.2756 + 0.1334) = 0.6319$

Summary Table A					
Methionine	Lysine				Total
	0	0.05	0.10	0.15	
0	5.05	5.38	4.81	4.31	19.55
0.025	4.57	4.82	5.36	5.33	20.08
0.050	4.97	5.11	5.38	5.38	20.84
Total	14.59	15.31	15.55	15.02	60.47

TABLE 12.10.1—(Continued)

Computations (continued):

6. Entries are sums of 2 levels of protein; $5.05 = 2.08 + 2.97$, etc.
7. Total in A: $(5.05^2 + \dots + 5.38^2)/4 - C = 0.3496$
8. Lysine, L: $(14.59^2 + \dots + 15.02^2)/12 - C = 0.0427$
9. Methionine, M: $(19.55^2 + 20.08^2 + 20.84^2)/16 - C = 0.0526$
10. LM: $0.3496 - (0.0427 + 0.0526) = 0.2543$

SUMMARY TABLE B

Methionine	Protein		Total
	12	14	
0	8.95	10.60	19.55
0.025	9.59	10.49	20.08
0.050	9.16	11.68	20.84
Total	27.70	32.77	60.47

Computations (continued):

11. Entries are sums of 4 levels of lysine; $8.95 = 2.08 + 2.30 + 2.35 + 2.22$, etc.
12. Total in B: $(8.95^2 + \dots + 11.68^2)/8 - C = 0.6702$
13. Protein, P: $(27.70^2 + 32.77^2)/24 - C = 0.5355$
14. MP: $0.6702 - (0.5355 + 0.0526) = 0.0821$

Summary Table C

Protein	Lysine				Total
	0	0.05	0.10	0.15	
12	6.22	6.62	7.63	7.23	27.70
14	8.37	8.69	7.92	7.79	32.77
Total	14.59	15.31	15.55	15.02	60.47

Computations (continued):

15. Entries are sums of 3 levels of methionine; $6.22 = 2.08 + 2.08 + 2.06$, etc.
16. Total in C: $(6.22^2 + \dots + 7.79^2)/6 - C = 0.8181$
17. LP: $0.8181 - (0.5355 + 0.0427) = 0.2399$
18. LMP: $1.2756 - (0.0427 + 0.0526 + 0.5355 + 0.2543 + 0.0821 + 0.2399) = 0.0685$

$$L_L P_L = \frac{(6.55)^2}{(6)(2)(20)} = 0.1788,$$

$F = 0.1788/0.0275 = 6.50$, $P = 0.025$. This corresponds to the highly significant effect observed in table 12.9.2, where an interpretation was given.

Deducting $L_L P_L$ from the LP sum of squares in table 12.10.2, $0.2399 - 0.1788 = 0.0611$, shows that neither of the other two comparisons can be significant.

EXAMPLE 12.10.3—The investigator is often interested in estimates of differences rather than in tests of significance. Because of the LP interaction he might wish to estimate the effect of protein with no lysine. Summary table C shows this mean difference:

TABLE 12.10 2
ANALYSIS OF VARIANCE OF 3-FACTOR PIG EXPERIMENT
RANDOMIZED BLOCKS DESIGN

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Replications	1	0.1334	
Lysine, $L(l = 4)$	3	0.0427	0.0142
Methionine, $M(m = 3)$	2	0.0526	0.0263
Protein, $P(p = 2)$	1	0.5355	0.5355**
LM	6	0.2543	0.0424
LP	3	0.2399	0.0800
MP	2	0.0821	0.0410
LMP	6	0.0685	0.0114
Error ($r = 2$)	23	0.6319	0.0275

$(8.37 - 6.22)/6 = 0.36$ lb/day (The justification for using all levels of methionine is that there is little evidence of either main effect or interaction with protein.) The standard error of the mean difference is $\pm \sqrt{(2)(0.0275)/6} = 0.096$ lb./day. Verify that the 95% interval is from 0.16 to 0.56 lb./day.

12.11—Expected values of mean squares. In the analysis of variance of a factorial experiment, the expected values of the mean squares for main effects and interactions can be expressed in terms of components of variance that are part of the mathematical model underlying the analysis. These formulas have two principal uses. They show how to obtain unbiased estimates of error for the comparisons that are of interest. In studies of variability they provide estimates of the contributions made by different sources to the variance of a measurement.

Consider a two-factor $A \times B$ experiment in a completely randomized design, with a levels of A , b levels of B , and n replications. The observed value for the k th replication of the i th level of A and the j th level of B is

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (12.11.1)$$

where $i = 1 \dots a$, $j = 1 \dots b$, $k = 1 \dots n$. (If the plan is in randomized blocks or a Latin square, further parameters are needed to specify block, row, or column effects.)

The parameters α_i and β_j , representing main effects, may be fixed or random. If either A or B is random, the corresponding α_i or β_j are assumed drawn from an infinite population with mean zero, variance σ_A^2 or σ_B^2 . The $(\alpha\beta)_{ij}$ are the two-factor interaction effects. They are random if either A or B is random, with mean 0, variance σ_{AB}^2 . As usual, the ε_{ijk} have mean 0, variance σ^2 .

Before working out the expected value of the mean square for A , we must be clear about the meaning of main effects. The relevant and useful way of defining the main effect of A , and consequently the expected value of its mean square, depends on whether the other factor B is fixed or random.

To illustrate the distinction, let A represent 2 fertilizers and B 2 fields. Experimental errors ε are assumed negligible, and results are as follows:

	Fertilizer		$a_2 - a_1$
	a_1	a_2	
Field 1	10	17	+7
Field 2	18	13	-5
Mean	14	15	+1

When B is fixed, our question is: What is the average difference between a_2 and a_1 over these two fields? The answer is that a_2 is superior by 1 unit (15 - 14). The answer is *exact*, since experimental errors are negligible in this example. But if B is random, the question becomes: What can be inferred about the average difference between a_2 and a_1 over a population of fields of which these two fields are a random sample? The difference ($a_2 - a_1$) is +7 in field 1 and -5 in field 2, with mean 1 as before. The estimate is no longer exact, but has a standard error (with 1 d.f.), which may be computed as $\sqrt{\{7 - (-5)\}^2/4} = \pm 6$. Note that this standard error is derived from the AB interaction, this interaction being, in fact, $\{7 - (-5)\}/2 = 6$.

To sum up, the numerical estimates of the main effects of A are the same whether B is fixed or random, but the population parameters being estimated are not the same, and hence different standard errors are required in the two cases.

From equation 12.11.1 the sample mean for the i th level of A is

$$\bar{X}_{i..} = \mu + \alpha_i + \bar{\beta} + (\bar{\alpha\beta})_{i.} + \bar{\varepsilon}_{i..} \quad (12.11.2)$$

where $\bar{\beta} = (\beta_1 + \dots + \beta_b)/b$, $(\bar{\alpha\beta})_{i.} = \{(\alpha\beta)_{i1} + \dots + (\alpha\beta)_{ib}\}/b$ and $\bar{\varepsilon}_{i..}$ is the average of nb independent values of ε .

When B is fixed, the true main effects of A are the differences of the quantities $\{\alpha_i + (\bar{\alpha\beta})_{i.}\}$ from level to level of A . In this case it is customary, for simplicity of notation, to redefine the parameter α_i as $\alpha'_i = \alpha_i + (\bar{\alpha\beta})_{i.}$ Thus with B fixed, it follows from equation 12.11.2 that

$$\bar{X}_{i..} - \bar{X}... = \alpha'_i - \bar{\alpha}' + \bar{\varepsilon}_{i..} - \bar{\varepsilon}... \quad (12.11.3)$$

From this relation the expected value of the mean square for A is easily shown to be

$$E(A) = E \left[\frac{nb \sum (\bar{X}_{i..} - \bar{X}...)^2}{a - 1} \right] = \frac{nb \sum (\alpha'_i - \bar{\alpha}')^2}{a - 1} + \sigma^2 \quad (12.11.4)$$

The quantity $\Sigma(\alpha_i' - \bar{\alpha}')^2/(a-1)$ is the quantity previously denoted by κ_A^2 .

If A is random and B is fixed, repeated sampling involves drawing a fresh set of a levels of the factor A in each experiment, retaining the same set of b levels of B . In finding $E(A)$ we average first over samples that happen to give the same set of levels of A , this being a common device in statistical theory. Formula 12.11.4 holds at this stage. When we average further over all sets of a levels of A that can be drawn from the population, κ_A^2 is an unbiased estimate of σ_A^2 , the population variance of α_i' . Hence, with A random and B fixed,

$$E(A) = nb\sigma_A^2 + \sigma^2$$

Now consider B random and revert to equation 12.11.2.

$$\bar{X}_{i..} = \mu + \alpha_i + \bar{\beta} + (\bar{\alpha\beta})_{i.} + \bar{\epsilon}_{i..} \quad (12.11.2)$$

In each new sample we draw fresh values of β_j and of $(\alpha\beta)_{ij}$ so that β and $(\alpha\beta)_{i.}$ change from sample to sample. Since, however, the population means of $\bar{\beta}$, $(\bar{\alpha\beta})_{i.}$ and $\bar{\epsilon}_{i..}$ are all zero, the population mean of $\bar{X}_{i..}$ is $\mu + \alpha_i$. Consequently, the population variance of the main effects of A is defined as $\kappa_A^2 = \Sigma(\alpha_i - \bar{\alpha})^2/(a-1)$ if A is fixed, or as the variance σ_A^2 of the α 's if A is random. But since

$$\bar{X}_{i..} - \bar{X}... = \alpha_i - \bar{\alpha} + (\bar{\alpha\beta})_{i.} - (\bar{\alpha\beta})_{..} + \bar{\epsilon}_{i..} - \bar{\epsilon}...,$$

the expected value of the mean square of A now involves σ_{AB}^2 as well as σ^2 .

It follows that when B is random,

$$E(A) = nb\kappa_A^2 + n\sigma_{AB}^2 + \sigma^2 \quad (A \text{ fixed})$$

$$E(A) = nb\sigma_A^2 + n\sigma_{AB}^2 + \sigma^2 \quad (A \text{ random})$$

The preceding results are particular cases of a more general formula. If the population of levels of B is finite, containing B' levels of which b are chosen at random for the experiment,

$$E(A) = nb\sigma_A^2 + n\left(\frac{B' - b}{B'}\right)\sigma_{AB}^2 + \sigma^2$$

This case occurs, for instance, if a combine of B' factories or cotton growers carries out experiments in a random sample of b factories or fields. If $b = B'$ the term in σ_{AB}^2 vanishes and we regard factor B as fixed. As B' tends to infinity, the coefficient of σ_{AB}^2 tends to n , factor B being random. If A is fixed, σ_A^2 becomes κ_A^2 .

The AB mean square is derived from the sum of squares of the terms $(X_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}...)$. From the model, this term is

$$(\alpha\beta)_{ij} - (\bar{\alpha\beta})_{i.} - (\bar{\alpha\beta})_{.j} + (\bar{\alpha\beta})_{..} + \bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..} - \bar{\epsilon}_{.j.} + \bar{\epsilon}...$$

Unless both A and B are fixed, the interaction term in the above is a random variable from sample to sample, giving

$$E(AB) = n\sigma_{AB}^2 + \sigma^2$$

With both factors fixed, σ_{AB}^2 is replaced by κ_{AB}^2 . Table 12.11.1 summarizes this series of results.

TABLE 12.11.1
EXPECTED VALUES OF MEAN SQUARES IN A TWO-FACTOR EXPERIMENT
EXPECTED VALUE = PARAMETERS ESTIMATED

Mean Squares	Fixed Effects	Random Effects	Mixed Model A Fixed, B Random
A	$\sigma^2 + nb\kappa_A^2$	$\sigma^2 + n\sigma_{AB}^2 + nb\sigma_A^2$	$\sigma^2 + n\sigma_{AB}^2 + nb\kappa_A^2$
B	$\sigma^2 + na\kappa_B^2$	$\sigma^2 + n\sigma_{AB}^2 + na\sigma_B^2$	$\sigma^2 + na\sigma_B^2$
AB	$\sigma^2 + n\kappa_{AB}^2$	$\sigma^2 + n\sigma_{AB}^2$	$\sigma^2 + n\sigma_{AB}^2$
Error	σ^2	σ^2	σ^2

Note that when B is random and the main effects of A are 0 (κ_A^2 or $\sigma_A^2 = 0$), the mean square for A is an unbiased estimate of $\sigma^2 + n\sigma_{AB}^2$. It follows that the appropriate denominator or "error" for an F -test of the main effects of A is the AB Interactions mean square, as illustrated from our sample of two fields. When B is fixed, the appropriate denominator is the Error mean square in table 12.11.1.

General rules are available for factors A, B, C, D, \dots at levels a, b, c, d, \dots with n replications of each treatment combination. Any factors may be fixed or random. In presenting these rules, the symbol U denotes the factorial effect in whose mean square we are interested (for instance, the main effect of A , or the BC interaction, or the ACD interaction).

Rule 12.11.1. The expected value of the mean square for U contains a term in σ^2 and a term in σ_U^2 . It also contains a variance term for any interaction in which (i) *all* the letters in U appear, and (ii) *all* the other letters in the interaction represent random effects.

Rule 12.11.2. The coefficient of the term in σ^2 is 1. The coefficient of any other variance is n times the product of all letters a, b, c, \dots that do *not* appear in the set of capital letters A, B, C, \dots specifying the variance.

For example, consider the mean square for C in a three-way factorial. If A and B are both random,

$$E(C) = \sigma^2 + n\sigma_{ABC}^2 + nb\sigma_{AC}^2 + na\sigma_{BC}^2 + nab\sigma_C^2$$

If A is fixed but B is random, the terms in σ_{ABC}^2 and σ_{AC}^2 drop out by Rule 12.11.1, and we have

$$E(C) = \sigma^2 + na\sigma_{BC}^2 + nab\sigma_C^2$$

If A and B are both fixed, the expected value is

$$E(C) = \sigma^2 + nab\sigma_c^2$$

For main effects and interactions in which all factors are fixed, we have followed the practice of replacing σ^2 by κ^2 . Most writers use the symbol σ^2 in either case. Table 12.11.2 illustrates the rules for three factors.

TABLE 12.11.2
EXPECTED VALUES OF MEAN SQUARES IN A THREE-WAY FACTORIAL

Mean Squares	Expected Values	
	All Effects Fixed	All Effects Random
A	$\sigma^2 + nb\kappa_A^2$	$\sigma^2 + n\sigma_{ABC}^2 + n\sigma_{AB}^2 + nb\sigma_{AC}^2 + nb\kappa_A^2$
B	$\sigma^2 + na\kappa_B^2$	$\sigma^2 + n\sigma_{ABC}^2 + n\sigma_{AB}^2 + na\sigma_{BC}^2 + na\kappa_B^2$
C	$\sigma^2 + nab\kappa_C^2$	$\sigma^2 + n\sigma_{ABC}^2 + nb\sigma_{AC}^2 + na\sigma_{BC}^2 + nab\sigma_C^2$
AB	$\sigma^2 + n\kappa_{AB}^2$	$\sigma^2 + n\sigma_{ABC}^2 + n\sigma_{AB}^2$
AC	$\sigma^2 + nb\kappa_{AC}^2$	$\sigma^2 + n\sigma_{ABC}^2 + nb\sigma_{AC}^2$
BC	$\sigma^2 + na\kappa_{BC}^2$	$\sigma^2 + n\sigma_{ABC}^2 + na\sigma_{BC}^2$
ABC	$\sigma^2 + n\kappa_{ABC}^2$	$\sigma^2 + n\sigma_{ABC}^2$
Error	σ^2	σ^2

Mean Squares	A Fixed, B and C Random
A	$\sigma^2 + n\sigma_{ABC}^2 + n\sigma_{AB}^2 + nb\sigma_{AC}^2 + nb\kappa_A^2$
B	$\sigma^2 + na\sigma_{BC}^2 + na\kappa_B^2$
C	$\sigma^2 + na\sigma_{BC}^2 + nab\sigma_C^2$
AB	$\sigma^2 + n\sigma_{ABC}^2 + n\sigma_{AB}^2$
AC	$\sigma^2 + n\sigma_{ABC}^2 + nb\sigma_{AC}^2$
BC	$\sigma^2 + na\sigma_{BC}^2$
ABC	$\sigma^2 + n\sigma_{ABC}^2$
Error	σ^2

From these formulas, unbiased estimates of all the components of variance can be obtained as linear combinations of the mean squares in the analysis of variance. The null hypothesis that any component is 0 can be tested, though complications may arise. Consider the null hypothesis $\sigma_c^2 = 0$. Table 12.11.2 shows that if all effects are fixed, the appropriate denominator for testing the mean square for C is the ordinary Error mean square of the experiment. If A is fixed and B is random, the appropriate denominator is the BC mean square.

If all effects are random, no single mean square in the analysis of variance is an appropriate denominator for testing σ_c^2 (check with table 12.11.2). An approximate F -test is obtained as follows (11, 12). If $\sigma_c^2 = 0$, you may verify from table 12.11.2 that

$$E(C) = E(AC) + E(BC) - E(ABC)$$

while if σ_C^2 is large, $E(C)$ will exceed the right-hand side. A test criterion is

$$F' = \{(C) + \frac{1}{3}(ABC)\} / \{(AC) + \frac{1}{3}(BC)\}$$

where (C) denotes the mean square for C , and so on. The approximate degrees of freedom are

$$n_1 = \frac{\{(C) + (ABC)\}^2}{\frac{(C)^2}{f_C} + \frac{(ABC)^2}{f_{ABC}}}$$

$$n_2 = \frac{\{(AC) + (BC)\}^2}{\frac{(AC)^2}{f_{AC}} + \frac{(BC)^2}{f_{BC}}}$$

12.12—The split-plot or nested design. It is often desirable to get precise information on one factor and on the interaction of this factor with a second, but to forego such precision on the second factor. For example, three sources of vitamin might be compared by trying them on three males of the same litter, replicating the experiment on 20 litters. This would be a randomized blocks design with high precision, providing 38 degrees of freedom for error. Superimposed on this could be some experiment with the litters as units. Four types of housing could be tried, one litter to each type, thus allowing 5 replications with 12 degrees of freedom for error. The main treatments (housings) would not be compared as accurately as the sub-treatments (sources of vitamin) for two reasons; less replication is provided, and litter differences are included in the error for evaluating the housing effects. Nevertheless, some information about housing may be got at little extra expense, and any interaction between housing and vitamin will be accurately evaluated.

In experiments on varieties or fertilizers on small plots, cultural practices with large machines may be tried on whole groups of the smaller plots, each group containing all the varieties. (Irrigation is one practice that demands large areas per treatment.) The series of cultural practices is usually replicated only a small number of times but the varieties are repeated on every cultural plot. Experiments of this type are called *split-plot*, the cultural *main plot* being split into smaller varietal *sub-plots*.

This design is also common in industrial research. Comparisons among relatively large machines, or comparisons of different conditions of temperature and humidity under which machines work, are *main plot* treatments, while adjustments internal to the machines are *sub-plot* treatments. Since the word *plot* is inappropriate in such applications, the designs are often called *nested*, in the sense of section 10.16.

The essential feature of the split-plot experiment is that the sub-plot treatments are not randomized over the whole large block but only over the main plots. Randomization of the sub-treatments is newly done in each main plot and the main treatments are randomized in the large blocks.

BLOCK I	Ladak	D
		A
		C
		B
	Ranger	B
		A
		D
		C
	Cossack	A
		C
		D
		B

BLOCK II	Ranger	C
		D
		B
		A
	Cossack	B
		A
		D
		C
	Ladak	A
		C
		B
		D

FIG. 12.12.1—First 2 blocks of split-plot experiment on alfalfa, illustrating random arrangement of main and sub-plots.

A consequence is that the experimental error for sub-treatments is different (characteristically smaller) than that for main treatments.

Figure 12.12.1 shows the field layout of a split-plot design with three varieties of alfalfa, the sub-treatments being four dates of final cutting (13). The first two harvests were common to all plots, the second on July 27, 1943. The third harvests were: *A*, none; *B*, September 1; *C*, September 20; *D*, October 7. Yields in 1944 are recorded in table 12.12.1. Such an experiment is, of course, not evaluated by a single season's yields; statistical methods for perennial crops are discussed in section 12.14.

In the analysis of variance the main plot analysis is that of randomized blocks with three varieties replicated in six blocks. The sub-plot analysis contains the sums of squares for dates of cutting, for the date \times variety interactions, and for the sub-plot error, found by subtraction as shown at the foot of table 12.12.2.

The significant differences among dates of cutting were not unexpected, nor were the smaller yields following *B* and *C*. The last harvest should be either early enough to allow renewed growth and restoration of the consequent depletion of root reserves, or so late that no growth and depletion will ensue. The surprising features of the experiment were two; the yield following *C* being greater than *B*, since late September is usually considered a poor time to cut alfalfa in Iowa; and the absence of interaction between date and variety—Ladak is slow to renew growth after cutting and might have reacted differently from the other varieties.

In order to justify this analysis we need to study the model. In randomized blocks, the model for the split-plot or nested experiment is

$$X_{ijk} = \mu + M_i + B_j + \varepsilon_{ij} + T_k + (MT)_{ik} + \delta_{ijk}$$

$$i = 1 \quad m, j = 1 \dots b, k = 1 \dots t, \varepsilon_{ij} = \mathcal{N}(0, \sigma_M), \delta_{ijk} = \mathcal{N}(0, \sigma_I)$$

Here, *M* stands for main plot treatments, *B* for blocks, and *T* for sub-plot treatments.

TABLE 12.12.1
YIELDS OF THREE VARIETIES OF ALFALFA (TONS PER ACRE) IN 1944 FOLLOWING
FOUR DATES OF FINAL CUTTING IN 1943

Variety	Date	Blocks					
		1	2	3	4	5	6
Ladak	A	2.17	1.88	1.62	2.34	1.58	1.66
	B	1.58	1.26	1.22	1.59	1.25	0.94
	C	2.29	1.60	1.67	1.91	1.39	1.12
	D	2.23	2.01	1.82	2.10	1.66	1.10
	.	8.27	6.75	6.33	7.94	5.88	4.82
Cossack	A	2.33	2.01	1.70	1.78	1.42	1.35
	B	1.38	1.30	1.85	1.09	1.13	1.06
	C	1.86	1.70	1.81	1.54	1.67	0.88
	D	2.27	1.81	2.01	1.40	1.31	1.06
	.	7.84	6.82	7.37	5.81	5.53	4.35
Ranger	A	1.75	1.95	2.13	1.78	1.31	1.30
	B	1.52	1.47	1.80	1.37	1.01	1.31
	C	1.55	1.61	1.82	1.56	1.23	1.13
	D	1.56	1.72	1.99	1.55	1.51	1.33
	.	6.38	6.75	7.74	6.26	5.06	5.07
Total		22.49	20.32	21.44	20.01	16.47	14.24

Variety	Date of Cutting				Total
	A	B	C	D	
Ladak	11.25	7.84	9.98	10.92	39.99
Cossack	10.59	7.81	9.46	9.86	37.72
Ranger	10.22	8.48	8.90	9.66	37.26
Total	32.06	24.13	28.34	30.44	114.97
Mean (tons per acre)	1.78	1.34	1.57	1.69	

The symbols i, j identify the main plot, while k identifies the sub-plot within the main plot. The two components of error, ε_{ij} and δ_{ijk} , are needed to make the model realistic: the sub-plots in one main plot often yield consistently higher than those in another, and ε_{ij} represents this difference. From the model, the error of the mean difference between two main plot treatments, say M_1 and M_2 , is

$$\bar{\varepsilon}_1 - \bar{\varepsilon}_2 + \bar{\delta}_1 - \bar{\delta}_2$$

The $\bar{\varepsilon}$'s are averages over b values, the $\bar{\delta}$'s over bt values. Consequently, the variance of the mean difference is

TABLE 12.12.2
ANALYSIS OF VARIANCE OF SPLIT-PLOT EXPERIMENT ON ALFALFA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Main plots:			
Varieties	2	0.1781	0.0890
Blocks	5	4.1499	0.8300
Main plot error	10	1.3622	0.1362
Sub-plots:			
Dates of cutting	3	1.9625	0.6542**
Date \times variety	6	0.2105	0.0351
Sub-plot error	45	1.2586	0.0280

1. Correction: $C = (114.97)^2/72 = 183.5847$

2. Total: $(2.17)^2 + \dots + (1.33)^2 - C = 9.1218$

3. Main plots: $\frac{(8.27)^2 + \dots + (5.07)^2}{4} - C = 5.6902$

4. Varieties: $\frac{(39.99)^2 + \dots + (37.26)^2}{24} - C = 0.1781$

5. Blocks: $\frac{(22.49)^2 + \dots + (14.24)^2}{12} - C = 4.1499$

6. Main plot error: $5.6902 - (0.1781 + 4.1499) = 1.3622$

7. Sub-classes in variety-date table: $\frac{(11.25)^2 + \dots + (9.66)^2}{6} - C = 2.3511$

8. Dates: $\frac{(32.06)^2 + \dots + (30.44)^2}{18} - C = 1.9625$

9. Date \times variety: $2.3511 - (0.1781 + 1.9625) = 0.2105$

10. Sub-plot error: $9.1218 - (5.6902 + 1.9625 + 0.2105) = 1.2586$

$$2\left(\frac{\sigma_M^2}{b} + \frac{\sigma_I^2}{bt}\right) = \frac{2}{bt}(\sigma_I^2 + t\sigma_M^2)$$

In the analysis of variance, the main plot Error mean square estimates $(\sigma_I^2 + t\sigma_M^2)$.

Consider now the difference $X_{ij1} - X_{ij2}$ between two sub-plots that are in the same main plot. According to the model,

$$X_{ij1} - X_{ij2} = T_1 - T_2 + (MT)_{i1} - (MT)_{i2} + \delta_{ij1} - \delta_{ij2}$$

The error now involves only the δ 's. Consequently, for any comparison among treatments that is made entirely within main plots, the basic error variance is σ_I^2 , estimated by the sub-plot Error mean square. Such comparisons include (i) the main effects of sub-plot treatments, (ii) interactions between main-plot and sub-plot treatments, and (iii) comparisons

between sub-plot treatments for a single main-plot treatment (e.g., between dates for Ladak).

In some experiments it is feasible to use either the split-plot design or ordinary randomized blocks in which the *mt* treatment combinations are randomized within each block. On the average, the two arrangements have the same overall accuracy. Relative to randomized blocks, the split-plot design gives reduced accuracy on the main-plot treatments and increased accuracy on sub-plot treatments and interactions. In some industrial experiments conducted as split-plots, the investigator apparently did not realize the implications of the split-plot arrangement and analyzed the design as if it were in randomized blocks. The consequences were to assign too low errors to main-plot treatments and too high errors to sub-plot treatments.

TABLE 12.12.3
PRESENTATION OF TREATMENT MEANS (TONS PER ACRE) AND STANDARD ERRORS

Variety	Date of Cutting ($\pm \sqrt{E_b/b} = \pm 0.0683$)				Means
	A	B	C	D	
Ladak	1.875	1.307	1.664	1.820	1.667 ($\pm \sqrt{E_a/tb} =$ ± 0.0753)
Cossack	1.765	1.302	1.577	1.644	
Ranger	1.704	1.414	1.484	1.610	
Means	1.781	1.341	1.575	1.691	
		$(\pm \sqrt{E_b/mb} = \pm 0.0394)$			

Care is required in the use of the correct standard errors for comparisons among treatment means. Table 12.12.3 shows the treatment means and *s.e.*'s for the alfalfa experiment, where $E_a = 0.1362$ and $E_b = 0.0280$ denote the main- and sub-plot Error mean squares. The *s.e.* ± 0.0683 , which is derived from E_b , is the basis for computing the *s.e.* for comparisons that are part of the Variety-Date interactions and for comparisons among dates for a single variety or a group of the varieties. The *s.e.* ± 0.0753 for varietal means is derived from E_a . Some comparisons, for example those among varieties for Date A, require a standard error that involves both E_a and E_b , as described in (8).

Formally, the sub-plot error *S.S.* (45 *df.*) is the combined *S.S.* for the *BT* interactions (15 *df.*) and the *BMT* interactions (30 *df.*). Often, it is more realistic to regard Blocks as a random component rather than as a fixed component. In this case, the error for testing *T* is the *BT* mean square, while that for testing *MT* is the *BMT* mean square, if the two mean squares appear to differ.

Experimenters sometimes split the sub-plots and even the sub-sub-plots. The statistical methods are a natural extension of those given here. If T_1, T_2, T_3 denote the sets of treatments at three levels, the set T_1 are tested against the main-plot Error mean square, T_2 and the $T_1 T_2$ interac-

tions against the sub-plot error, and T_3 , T_1T_3 , T_2T_3 , and $T_1T_2T_3$ against the sub-sub-plot error. For missing data see (8, 14).

EXAMPLE 12.12.1—A split-split-plot experiment on corn was conducted to try 3 rates of planting (stands) with 3 levels of fertilizer in irrigated and non-irrigated plots (21). The design was randomized blocks with 4 replications. The main plots carried the irrigation treatments. On each there were sub-plots with 3 stands, 10,000, 13,000, and 16,000 plants per acre. Finally, each sub-plot was divided into 3 parts respectively fertilized with 60, 120, and 180 pounds of nitrogen. The yields are in bushels per acre. Calculate the analysis of variance.

			Blocks			
			1	2	3	4
			90	83	85	86
Not Irrigated	Stand 1	Fertilizer 1	95	80	88	78
		2	107	95	88	89
		3				
	2	1	92	98	112	79
		2	89	98	104	86
		3	92	106	91	87
	3	1	81	74	82	85
		2	92	81	78	89
		3	93	74	94	83
Irrigated	1	1	80	102	60	73
		2	87	109	104	114
		3	100	105	114	114
	2	1	121	99	90	109
		2	110	94	118	131
		3	119	123	113	126
	3	1	78	136	119	116
		2	98	133	122	136
		3	122	132	136	133

Source of Variation	Degrees of Freedom	Mean Square
Main Plots:		
Blocks	3	
Irrigation, I	1	8,277.56
Error (a)	3	470.59
Sub-plots:		
Stand, S	2	879.18
IS	2	1,373.51*
Error (b)	12	232.33
Sub-sub-plots:		
Fertilizer, F	2	988.72
IF	2	476.72**
SF	4	76.22
ISF	4	58.68
Error (c)	36	86.36

EXAMPLE 12.12.2—Attention is attracted to the two significant interactions, IS and IF . Now, ISF is less than error. This means that the IS interaction is much the same at all levels of F ; or, alternatively, that the IF interaction is similar at all levels of S . Hence, each 2-way table gives information.

	F_1	F_2	F_3	S_1	S_2	S_3
Not Irrigated	1,047	1,058	1,099	1,064	1,134	1,006
Irrigated	1,183	1,356	1,437	1,162	1,353	1,461

Neither fertilizer nor stand affected yield materially on the non-irrigated plots. With irrigation, the effect of each was pronounced. So it is necessary to examine separately the split-plot experiment on the irrigated plots. Verify the following mean squares:

Stand:			
Linear	1	3,725**	
Deviations	1	96	
Error (a)	6	316	
Fertilizer:			
Linear	1	2,688**	
Deviations	1	118	
SF	4	92	
Error (b)	18	137	

EXAMPLE 12.12.3—Notice that the planting and fertilizer rates were well chosen for the unirrigated plots, but on the irrigated plots they were too low to allow any evaluation of the optima. This suggests that irrigation should not be a factor in such experiments. But in order to compare costs and returns over a number of years, two experiments (one with and one without irrigation) should be randomly interplanted to control fertility differences.

12.13—Series of experiments. A series of experiments may extend over several places or over several years or both. In a number of countries in which the supply of food is deficient, such series have been undertaken in recent years on farmers' fields in order to estimate the amount by which the production of food grains can be increased by greater use of fertilizers.

Every series of experiments presents a unique problem for the experimenter and the statistician, both in planning and analysis. Good presentations of the difficulties involved are in (15, 16, 17, 18), with illustrations of the analysis. The methods given in this book should enable the reader to follow the references cited. Only a brief introduction to the analysis for experiments conducted at a number of places will be given here.

We suppose that the experiments are all of the same size and structure, and that the places can be regarded as a random sample of the region about which inferences are to be made. For many reasons, a strictly random sample of places is difficult to achieve in practice: insofar as the sample is unrepresentative, inferences drawn from the analysis are vulnerable to bias.

In the simplest case, the important terms in a combined analysis of variance are:

Treatments
 Treatments \times Places
 Pooled experimental errors

The Treatments \times Places mean square is tested against the pooled error (average of the Error mean squares in the individual experiments). If F is materially greater than 1, indicating that treatment effects change from place to place, the Treatments mean square is tested against the Treatments \times Places mean square, which becomes the basic error term for drawing conclusions about the average effects of treatments over the region.

Two complications occur. The experimental error variances often differ from place to place. This can be checked by Bartlett's test for homogeneity of variance. If variances are heterogeneous, the F -test of the Treatments \times Places interactions is not strictly valid, but an adjusted form of the test serves as an adequate approximation (15, 17). If comparisons are being made over a subset of the places, as suggested later, the pooled error for these places should be used instead of the overall pooled error.

Secondly, the Treatments \times Places interactions may not be homogeneous, especially in a factorial experiment. Some factors may give stable responses from place to place, while others are more erratic in their performance. If the Treatments mean square has been subdivided into sets of comparisons, the Interactions mean square for each set should be computed and tested separately.

The preceding approach is appropriate where the objective is to reach a single set of conclusions that apply to the whole region. Sometimes there is reason to expect that the relative performances of the treatments will vary with the soil type, with climatic conditions within the region, or with other characteristics of the places. The series may have been planned so as to examine such differences, leading perhaps to different recommendations for different parts of the region. In the analysis, the places then subdivide into a number of sets. The Treatments \times Places interactions are separated into

Treatments \times Sets
 Treatments \times Places within sets

If the Treatments \times Sets mean square is substantially larger than Treatments \times Places within sets, it is usually advisable to examine the results separately for each set.

The following examples illustrate the preliminary steps in the analysis of one series of experiments.

EXAMPLE 12 13.1—The following data illustrate a series of experiments over five places (21). Four treated lots of 100 Mukden soybean seeds, together with one lot untreated, were planted in 5 randomized blocks at each participating station. The total numbers of emerging plants (from 500 seeds) are shown for the 5 locations. Also shown are the analyses of variance at the several stations

NUMBER OF EMERGING PLANTS (500 SEEDS) IN FIVE PLOTS COOPERATIVE SEED
TREATMENT TRIALS WITH MUKDEN SOYBEANS, 1943

Location	Untreated	Arasan	Spergon	Semesan, Jr.	Fermate	Total
Michigan	360	356	362	350	373	1,801
Minnesota	302	354	349	332	332	1,669
Wisconsin	408	407	391	391	409	2,006
Virginia	244	267	293	235	278	1,317
Rhode Island	373	387	406	394	375	1,935
Total	1,687	1,771	1,801	1,702	1,767	8,722

Mean Squares From Original Analyses of Variance

Source of Variation	Degrees of Freedom	Location				
		Michigan	Minnesota	Wisconsin	Virginia	Rhode Island
Treatments	4	14.44	82.84*	17.44	114.26*	37.50
Blocks	4	185.14	54.64	5.64	70.76	4.80
Error	16	42.29	26.67	30.64	26.34	13.05

Test the hypothesis of homogeneity of error variance. Ans. Corrected $\chi^2 = 5.22$, $df = 4$.

EXAMPLE 12.13.2—For the entire soybean data, analyze the variance as follows:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Treatments	4	380.29	95.07
Locations	4	11,852.61	2,963.15
Interaction	16	685.63	42.85
Blocks in Locations	20	1,283.92	...
Experimental Error	80	2,223.68	27.80

Blocks and Experimental Error are pooled values from the analyses of the five places

EXAMPLE 12.13.3—Isolate the sum of squares for the planned comparison, Untreated vs. Average of the four Treatments. Ans. 171.70, $F = 4.01$, $F_{05} = 4.49$

12.14—Experiments with perennial crops. When a perennial crop is investigated over a number of years, the yields from the same plot in successive years are usually correlated. The experimental error in one season is not independent of that in another season.

In comparing the overall yields of the treatments, this difficulty is overcome by first finding for each plot the total yield over all years. These totals are analyzed by the method appropriate to the design that was used. This method provides a valid error for testing the overall treatment effects.

For illustration, the data in table 12.14.1 are taken from an experiment by Haber (19) to compare the effects of various cutting treatments on asparagus. Planting was in 1927 and cutting began in 1929. One plot in each block was cut until June 1 in each year, others to June 15, July 1, and July 15. The yields are for the four succeeding years, 1930, 1931,

1932, and 1933. The yields are the weights cut to June 1 in every plot, irrespective of later cuttings in some of them. This weight is a measure of vigor, and the objective is to compare the relative effectiveness of the different harvesting plans.

A glance at the four-year totals (5,706; 5,166; 4,653; 3,075) leaves little doubt that prolonged cutting decreased the vigor. The cutting totals were separated into linear, quadratic, and cubic components of the regres-

TABLE 12.14.1
WEIGHT (OUNCES) OF ASPARAGUS CUT BEFORE JUNE 1 FROM PLOTS WITH
VARIOUS CUTTING TREATMENTS

Blocks	Year	Cutting Ceased				Total
		June 1	June 15	July 1	July 15	
1	1930	230	212	183	148	773
	1931	324	415	320	246	1,305
	1932	512	584	456	304	1,856
	1933	399	386	255	144	1,184
		<u>1,465</u>	<u>1,597</u>	<u>1,214</u>	<u>842</u>	<u>5,118</u>
2	1930	216	190	186	126	718
	1931	317	296	295	201	1,109
	1932	448	471	387	289	1,595
	1933	361	280	187	83	911
		<u>1,342</u>	<u>1,237</u>	<u>1,055</u>	<u>699</u>	<u>4,333</u>
3	1930	219	151	177	107	654
	1931	357	278	298	192	1,125
	1932	496	399	427	271	1,593
	1933	344	254	239	90	927
		<u>1,416</u>	<u>1,082</u>	<u>1,141</u>	<u>660</u>	<u>4,299</u>
4	1930	200	150	209	168	727
	1931	362	336	328	226	1,252
	1932	540	485	462	312	1,799
	1933	381	279	244	168	1,072
		<u>1,483</u>	<u>1,250</u>	<u>1,243</u>	<u>874</u>	<u>4,850</u>
Total		5,706	5,166	4,653	3,075	18,600
		Degrees of Freedom	Sum of Squares		Mean Square	
Blocks		3	30,170			
Cuttings		(3)	(241,377)			
Linear		1			220,815**	
Quadratic		1			16,835*	
Cubic		1			3,727	
Error		9			2,429 -	

sion on duration of cutting. The significant quadratic component indicates that the yields fall off more and more rapidly as the severity of cutting increases.

Such experiments also contain information about the constancy of treatment differences from year to year, as indicated by the Treatments \times Years interactions. Often it is useful to compute on each plot the linear regression of yield on years, multiplying the yields in the four years by $-3, -1, +1, +3$ and adding. These linear regressions (with an appropriate divisor) measure the average rate of improvement of yield from year to year. An analysis of the linear regressions for the asparagus data appears in table 12.14.2. From the totals for each treatment it is evident that the improvement in yield per year is greatest for the June 1 cutting, and declines steadily with increased severity of cutting, the July 15 cutting showing only a modest total, 119.

TABLE 12.14.2
ANALYSIS OF THE LINEAR REGRESSION OF YIELD ON YEARS

Blocks	Cutting Ceased				Total
	June 1	June 15	July 1	July 15	
1	695*	691	352	46	1,784
2	566	445	95	-41	1,065
3	514	430	315	28	1,287
4	721	536	239	86	1,582
Total	2,496	2,102	1,001	119	5,718

	Degrees of Freedom	Sum of Squares	Mean Square
Blocks	3	3,776	
Cuttings.	(3)	43,633	14,544**
Linear	1	42,354**	
Quadratic	1	744	
Cubic	1	536	
Error	9	2,236	248

* $695 = 3(399) + 512 - 324 - 3(230)$, from table 12.14.1

In the analysis of variance of these linear regression terms, the sum of squares between cuttings has been subdivided into its linear, quadratic, and cubic regression on duration. Only the linear term was strongly significant. Evidently, each additional two weeks of cutting produced about the same decrease in the annual rate of improvement of yield.

In this analysis of variance an extra divisor $20 = 3^2 + 1^2 + 1^2 + 3^2$ was applied to each sum of squares, in order that the mean squares refer to a single observation. Can you explain why the Error mean square, 248, is so much smaller than the Error mean square for the four-year totals, 2,429? Features of this experiment have been discussed by Snedecor and Haber (19, 20).

REFERENCES

1. R. A. FISHER. *The Design of Experiments*. Oliver and Boyd, Edinburgh (1935-1951).
2. F. YATES. *J. R. Statist. Soc. Supp.*, 2:210 (1935).
3. Southern Cooperative Series Bulletin No. 10, p. 114 (1951).
4. Rothamsted Experimental Station Report: 218 (1937).
5. A. R. SAUNDERS. Union of South Africa Dept. of Agriculture and Forestry Sci. Bul. No. 200 (1939).
6. G. W. SNEDECOR. *Proc. Int. Statist. Conferences*, 3:440 (1947).
7. O. L. DAVIES (ed.). *Design and Analysis of Industrial Experiments*, 2nd ed. Oliver and Boyd, Edinburgh (1956).
8. W. G. COCHRAN and G. M. COX. *Experimental Designs*. Wiley, New York (1957).
9. Iowa Agricultural Experiment Station, Animal Husbandry Swine Nutrition Experiment No. 577 (1952).
10. F. YATES. "The Design and Analysis of Factorial Experiments." *Commonwealth Bureau of Soil Science Tech. Comm.* 35 (1937).
11. W. G. COCHRAN. *Biometrics*, 7:17 (1951).
12. F. E. SATTERTHWAIT. *Biometrics Bul.*, 2:110 (1946).
13. C. P. WILSIE. Iowa State College Agricultural Experiment Station (1944).
14. R. L. ANDERSON. *Biometrics Bul.*, 2:41 (1946).
15. F. YATES and W. G. COCHRAN. *J. Agric. Sci.*, 28:556 (1938).
16. O. KEMPTHORNE. *The Design and Analysis of Experiments*. Wiley, New York (1952).
17. W. G. COCHRAN. *Biometrics*, 10:101 (1954).
18. F. YATES, S. LIPTON, P. SINHA, and K. P. DASGUPTA. *Emp. J. Exp. Agric.*, 27:263 (1959).
19. E. S. HABER and G. W. SNEDECOR. *Amer. Soc. Hort. Sci.*, 48:481 (1946).
20. G. W. SNEDECOR and E. S. HABER. *Biometrics Bul.*, 2:61 (1946).
21. R. H. PORTER. Cooperative Soybean Seed Treatment Trials, Iowa State University Seed Laboratory (1943).

Multiple regression

13.1—Introduction. The regression of Y on a single independent variable (chapter 6) is often inadequate. Two or more X 's may be available to give additional information about Y by means of a multiple regression on the X 's. Among the principal uses of multiple regression are:

(1) Constructing an equation in the X 's that gives the best prediction of the values of Y .

(2) When there are many X 's, finding the subset that gives the best linear prediction equation. In predicting future weather conditions at an airport, there may be as many as 50 available X -variables, which measure different aspects of the present weather pattern at neighboring weather stations. A prediction equation with 50 variables is unwieldy, and is unwise if many of the X -variables contribute nothing to improved accuracy in the prediction. An equation based on the best three or four variables might be a wise choice.

(3) In some studies the objective is not prediction, but instead to discover which variables are related to Y , and, if possible, to rate the variables in order of their importance.

Multiple regression is a complex subject. The calculations become lengthy when there are numerous X -variables, and it is hard to avoid mistakes in computation. Standard electronic computer programs, now becoming more readily available, are a major help. Equally important is an understanding of what a multiple regression equation means and what it does not mean. Fortunately, much can be learned about the basis of the computations and the pitfalls in interpretation by study of a regression on two X -variables, which will be considered in succeeding sections before proceeding to three or more X -variables.

13.2—Two independent variables. With only one X -variable, the sample values of Y and X could be plotted as in figures 6.2.1 and 6.4.1, which show both the regression line and the distributions of the individual values of Y about the line. But if Y depends partly on X_1 and partly on

X_2 for its value, solid geometry instead of plane is required. Any observation now involves three numbers—the values of Y , X_1 , and X_2 . The pair (X_1, X_2) can be represented by a point on graph paper. The values of Y corresponding to this point are on a vertical axis perpendicular to the graph paper. In the population these values of Y form a frequency distribution, so we must try to envisage a frequency distribution of Y on each vertical axis. Each frequency distribution has a mean—the mean value of Y for specified X_1, X_2 . The surface determined by these means is the *regression surface*. In this chapter the surface is a *plane*, since only linear regressions on X_1 and X_2 are being studied.

The population regression plane is written

$$Y_R = \alpha + \beta_1 X_1 + \beta_2 X_2,$$

where Y_R denotes the mean value of the frequency distribution of Y for specified X_1, X_2 . In mathematical notation, $Y_R = E(Y|X_1, X_2)$.

What does β_1 measure? Suppose that the value of X_1 increases by 1 unit, while the value of X_2 remains unchanged. Y_R becomes

$$Y_R' = \alpha + \beta_1 X_1 + \beta_1 + \beta_2 X_2 = Y_R + \beta_1$$

Thus, β_1 measures *the average or expected change in Y when X_1 increases by 1 unit, X_2 remaining unchanged*. For this reason β_1 is called the *partial regression coefficient* of Y on X_1 . Some writers use a more explanatory symbol $\beta_{Y_1 \cdot 2}$ for β_1 , the subscript $\cdot 2$ being a reminder that X_2 also appears in the regression equation.

For given X_1, X_2 , the individual values of Y vary about the regression plane in a normal distribution with mean 0 and variance σ^2 , sometimes denoted by $\sigma_{Y \cdot 12}^2$. Hence, the model is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \varepsilon = \mathcal{N}(0, \sigma) \quad (13.2.1)$$

Given a sample of n values of (Y, X_1, X_2) the sample regression—the prediction equation—is

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 \quad (13.2.2)$$

The values of a, b_1 , and b_2 are chosen so as to minimize $\Sigma(Y - \hat{Y})^2$, the sum of squares of the n differences between the actual and the predicted Y values. With our model, theory shows that the resulting estimates a, b_1, b_2 , and \hat{Y} are unbiased and have the smallest standard errors of any unbiased estimates that are linear expressions in the Y 's. The value of a is given by the equation

$$a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 \quad (13.2.3)$$

By substitution for a in (13.2.2) the fitted regression can be written

$$\hat{Y} = \bar{Y} + b_1 x_1 + b_2 x_2, \quad (13.2.4)$$

where $x_1 = X_1 - \bar{X}_1$, as usual.

The b 's satisfy the *normal equations*:

$$b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 = \Sigma x_1 y \quad (13.2.5)$$

$$b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 = \Sigma x_2 y \quad (13.2.6)$$

Solution of these equations by standard algebraic methods leads to the formulas:

$$b_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{D} \quad (13.2.7)$$

and

$$b_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{D}, \quad (13.2.8)$$

where

$$D = (\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2 \quad (13.2.9)$$

The illustration (table 13.2.1) is taken from an investigation (1) of the source from which corn plants in various Iowa soils obtain their phosphorus. The concentrations of inorganic (X_1) and organic (X_2) phosphorus in the soils were determined chemically. The phosphorus content Y of corn grown in the soils was also measured.

The familiar calculations under the table give the sample means and the sums of squares and products of deviations from the means. Substitution in (13.2.7) to (13.2.9) gives

$$D = (1,752.96)(3,155.78) - (1,085.61)^2 = 4,353,400$$

$$b_1 = \frac{(3,155.78)(3,231.48) - (1,085.61)(2,216.44)}{4,353,400} = 1.7898$$

$$b_2 = \frac{(1,752.96)(2,216.44) - (1,085.61)(3,231.48)}{4,353,400} = 0.0866$$

From (13.2.3), a is given by

$$a = 81.28 - (1.7898)(11.94) - (0.0866)(42.11) = 56.26$$

The multiple regression equation becomes

$$\hat{Y} = 56.26 + 1.7898X_1 + 0.0866X_2 \quad (13.2.10)$$

The meaning is this: For each additional part per million of inorganic phosphorus in the soil at the beginning of the growing season, the phosphorus in the corn increased by 1.7898 ppm, as against 0.0866 ppm for each additional ppm of organic phosphorus. The suggestion is that the inorganic phosphorus in the soil was the chief source of plant-available phosphorus. This deduction needs further consideration (sections 13.3 and 13.5).

TABLE 13.2.1
INORGANIC PHOSPHORUS X_1 , ORGANIC PHOSPHORUS X_2 , AND ESTIMATED PLANT-AVAILABLE
PHOSPHORUS Y IN 18 IOWA SOILS AT 20° C. (PARTS PER MILLION)

Soil Sample	X_1	X_2	Y	\hat{Y}	$Y - \hat{Y}$
1	0.4	53	64	61.6*	2.4*
2	0.4	23	60	59.0	1.0
3	3.1	19	71	63.4	7.6
4	0.6	34	61	60.3	0.7
5	4.7	24	54	66.7	-12.7
6	1.7	65	77	64.9	12.1
7	9.4	44	81	76.9	4.1
8	10.1	31	93	77.0	16.0
9	11.6	29	93	79.6	13.4
10	12.6	58	51	83.8	-32.8
11	10.9	37	76	79.0	-3.0
12	23.1	46	96	101.6	-5.6
13	23.1	50	77	101.9	-24.9
14	21.6	44	93	98.7	-5.7
15	23.1	56	95	102.4	-7.4
16	1.9	36	54	62.8	-8.8
17	26.8	58	168	109.2	58.8
18	29.9	51	99	114.2	-15.2
Sum	215.0	758	1,463	1,463.0	0.0
Mean	11.94	42.11	81.28		
<hr/>					
$\Sigma X_1^2 = 4,321.02$	$\Sigma X_1 X_2 = 10,139.50$		$\Sigma X_1 Y = 20,706.20$		
$C = 2,568.06$	$C = 9,053.89$		$C = 17,474.72$		
$\Sigma x_1^2 = 1,752.96$	$\Sigma x_1 x_2 = 1,085.61$		$\Sigma x_1 y = 3,231.48$		
$\Sigma X_2^2 = 35,076.00$	$\Sigma X_2 Y = 63,825.00$		$\Sigma Y^2 = 131,299.00$		
$C = 31,920.22$	$C = 61,608.56$		$C = 118,909.39$		
$\Sigma x_2^2 = 3,155.78$	$\Sigma x_2 y = 2,216.44$		$\Sigma y^2 = 12,389.61$		

* The number of significant digits retained in the preceding calculations will affect these columns by ± 0.1 or ± 0.2

From the fitted regression (equation 13.2.10), the predicted value \hat{Y} can be estimated for each soil sample in table 13.2.1. For example, for soil 1,

$$\hat{Y} = 56.26 + 1.7898(0.4) + 0.0866(53) = 61.6 \text{ ppm}$$

The observed value $Y = 64$ ppm deviates by $64 - 61.6 = +2.4$ ppm from the estimated regression value. The 18 values of \hat{Y} are recorded in table 13.2.1. The deviations $Y - \hat{Y}$ are in the final column; they measure the failure of the X 's to predict Y .

The investigator now has the opportunity to examine the deviations from regression. In part they might be associated with other variables not included in the study. Or some explanation might be found for certain

deviations—especially the larger ones. Such explanation might be a valuable finding of the analysis, providing clues for further experimentation, or it might lead to the rejection of one or more observations and to a recalculation of the regression. In the present example the results for soil 17 immediately strike the eye. This soil has much the highest value of Y , 168. Before the regression was calculated, this value might not seem necessarily out of line (though it should be verified from the records), because soil 17 has the second highest value of both types of soil phosphorus, which could account for the high plant phosphorus. But this soil also has the highest deviation $Y - \hat{Y} = +58.8$. A test of this deviation will be presented in section 13.5.

A check on the linearity of the regression is made by plotting two scatter diagrams. First, plot the deviations $Y - \hat{Y}$ against X_1 , then plot the same deviations against X_2 . If the regression is markedly non-linear in one of the X 's, a curve instead of a horizontal straight line should be detectable in the corresponding graph. For curved multiple regression, see example 15.5.1.

13.3—The deviations mean square and the F -test. In the multiple regression model, the deviations of the Y 's from the population regression plane have mean 0 and variance σ^2 . An unbiased estimate of σ^2 is $s^2 = \Sigma(Y - \hat{Y})^2 / (n - k)$, where n is the size of sample and k is the number of parameters that have been estimated in fitting the regression. In the example $n = 18$ with 3 parameters α, β_1, β_2 , giving $n - k = 15$.

The deviations sum of squares $\Sigma(Y - \hat{Y})^2$ can be computed in two ways. If the individual deviations have been tabulated as in the last column of table 13.2.1, their sum of squares is run up directly, giving $\Sigma(Y - \hat{Y})^2 = 6,414.5$.

In practice a quicker method, based on an algebraic identity, is used. From equation (13.2.4) we had

$$\hat{Y} = \bar{Y} + b_1x_1 + b_2x_2$$

Since the sample means of x_1 and x_2 are both zero, the sample mean of the fitted values \hat{Y} is \bar{Y} . Write $\hat{y} = \hat{Y} - \bar{Y}$ and $d = Y - \hat{Y}$, so that d represents the observed deviation of Y from the fitted regression at this point. It follows that

$$y = Y - \bar{Y} = (\hat{Y} - \bar{Y}) + (Y - \hat{Y}) = \hat{y} + d. \quad (13.3.1)$$

Two important results, proved later in this section, are, first,

$$\Sigma y^2 = \Sigma \hat{y}^2 + \Sigma d^2 \quad (13.3.2)$$

This result states that the sum of squares of deviations of the Y 's from their mean splits into two parts: (i) the sum of squares of deviations of the fitted values from their mean, and (ii) the sum of squares of deviations from the fitted values. The sum of squares $\Sigma \hat{y}^2$ is appropriately called "the sum of squares due to regression." In geometrical treatments of

multiple regression, the relation (equation 3.3.2) may be shown to be an extension of Pythagoras' theorem to more than two dimensions.

The second result, of more immediate interest, is:

$$\text{S.S. due to regression} = \Sigma \hat{y}^2 = b_1 \Sigma x_1 y + b_2 \Sigma x_2 y \quad (13.3.3)$$

Hence, the sum of squares of deviations from the regression may be obtained by subtracting from Σy^2 the sum of products of the b 's with the right sides of the corresponding normal equations. For the example we have,

$$\Sigma \hat{y}^2 = (1.7898)(3,231.48) + (0.0866)(2,216.44) = 5,975.6$$

The value of Σd^2 is then

$$\Sigma d^2 = \Sigma y^2 - \Sigma \hat{y}^2 = 12,389.6 - 5,975.6 = 6,414.0$$

Besides being quicker, this method is less subject to rounding errors than the direct method. Agreement of the two methods is an excellent check on the regression computations.

The mean square of the deviations is $6,414.0/15 = 427.6$, with 15 *d.f.* The corresponding standard error, $\sqrt{427.6} = 20.7$, provides a measure of how closely the regression fits the data. If the purpose is to find a more accurate method of predicting Y , the size of this standard error is of primary importance. For instance, if current methods of predicting some critical temperature can do this with a standard error of 3.2 degrees, while a multiple regression gives a standard error of 4.7 degrees, it is obvious that the regression is no improvement on the current methods, though it might, after further study, be useful in conjunction with the current methods.

Sometimes the object of the regression analysis is to understand why Y varies, where the X 's measure variables that are thought to influence Y through some causal mechanism. For instance, Y might represent the yields of a crop grown on the same field for a number of years under uniform husbandry, while the X 's measure aspects of weather or insect infestation that influence crop yields (2). In such cases, it is useful to compare the Deviations mean square, $\Sigma d^2/(n - k)$, with the original mean square of Y , namely $\Sigma y^2/(n - 1)$. In our example the Deviations mean square is 427.6, while the original mean square is $12,389.61/17 = 728.8$.

The ratio, $427.6/728.8 = 0.59$, estimates the fraction of the variance of Y that is *not* attributable to the multiple regression, while its complement, 0.41, estimates the fraction that is "explained" by the X -variables. Even if the regression coefficients are clearly statistically significant, it is not uncommon to find that the fraction of the variance of Y attributable to the regression is much less than 1/2. This indicates that most of the variation in Y must be due to variables not included in the regression.

In some studies the investigator is not at all confident initially that any of the X s are related to Y . In this event an F -test of the null hypothesis $\beta_1 = \beta_2 = 0$ is helpful. The test is made from the analysis of variance in

TABLE 13.3.1
ANALYSIS OF VARIANCE OF PHOSPHORUS DATA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	<i>F</i>
Regression	2	$\Sigma \hat{y}^2 = 5,975.6$	2 987.8	6.99**
Deviations	15	$\Sigma d^2 = 6,414.0$	427.6	
Total	17	$\Sigma y^2 = 12,389.6$	728.8	

table 13.3.1. *F* is the ratio of the mean square due to regression to the Deviations mean square.

The *F*-value, 6.99, with 2 and 15 *df.*, is significant at the 1% level.

By an extension of this analysis, tests of significance of the individual *b*'s can be made. We have (from table 13.2.1)

$$\Sigma x_1 y = 3,231.48 : \Sigma x_1^2 = 1,752.96 : \Sigma x_2 y = 2,216.44 : \Sigma x_2^2 = 3,155.78$$

If we had fitted a regression of *Y* on *X*₁ alone, the regression coefficient would be $b_{Y1} = 3,231.48/1,752.96 = 1.8434$. The reduction in sum of squares due to this regression would be $(\Sigma x_1 y)^2 / \Sigma x_1^2 = (3,231.48)^2 / (1,752.96) = 5,957.0$ with 1 *df.* When both *X*₁ and *X*₂ were included in the regression the reduction in sum of squares was 5,975.6, with 2 *df.* (table 13.3.1). The difference $5,975.6 - 5,957.0 = 18.6$, with 1 *df.*, measures the additional reduction due to the inclusion of *X*₂, given that *X*₁ is already present, or in other words the unique contribution of *X*₂ to the regression. The null hypothesis $\beta_2 = 0$ is tested by computing $F = 18.6/427.6 = 0.04$, with 1 and 15 *df.*, where 427.6 is the deviations mean square. The test is shown in table 13.3.2. Since *F* is small, the null hypothesis is not rejected.

Similarly, the null hypothesis $\beta_1 = 0$ is tested by finding the additional reduction in sum of squares due to the inclusion of *X*₁ in the regres-

TABLE 13.3.2
TEST OF EACH *X* AFTER THE EFFECT OF THE OTHER HAS BEEN REMOVED

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	<i>F</i>
<i>X</i> ₁ and <i>X</i> ₂	2	$\Sigma \hat{y}^2 = 5,975.6$		
<i>X</i> ₁ alone	1	$(\Sigma x_1 y)^2 / \Sigma x_1^2 = 5,957.0$		
<i>X</i> ₂ after <i>X</i> ₁	1	18.6	18.6	0.04
<i>X</i> ₁ and <i>X</i> ₂	2	$\Sigma \hat{y}^2 = 5,975.6$		
<i>X</i> ₂ alone	1	$(\Sigma x_2 y)^2 / \Sigma x_2^2 = 1,556.7$		
<i>X</i> ₁ after <i>X</i> ₂	1	4,418.9	4,418.9	10.30**
Deviations	15	6,414.0	427.6	

sion after X_2 has already been included (table 13.3.2). In this case $F = 10.30$ is significant at the 1% level.

This method of testing a partial regression coefficient may appear strange at first, but is very general. If $\beta_k = 0$ when there are k X -variables, this means that the true model contains only $X_1 \dots X_{k-1}$. We fit a regression on $X_1 \dots X_{k-1}$, obtaining the reduction in sum of squares, R_{k-1} . Then we fit a regression on $X_1 \dots X_k$, obtaining the reduction R_k . If $\beta_k = 0$, it can be proved that $(R_k - R_{k-1})$ is simply an estimate of σ^2 , so that $F = (R_k - R_{k-1})/s^2$ should be about 1. If, however, β_k is not zero, the inclusion of X_k improves the fit and $(R_k - R_{k-1})$ tends to become large, so that F tends to become large. Later, we shall see that the same test can be made as a t -test of b_k .

Incidentally, it is worth comparing $b_{Y1} = 1.8434$ with the value $b_1 = b_{Y1.2} = 1.7898$ obtained when X_2 is included in the regression. Two points are important. The value of the regression coefficient has changed. In multiple regression, the value of any regression coefficient depends on the other variables included in the regression. Statements made about the size of a regression coefficient are not unique, being conditional on these other variables. Secondly, in this case the change is small—this gives some assurance that this regression coefficient is stable. With X_2 , we have $b_{Y2} = 2,216.44/3,155.78 = 0.7023$, much larger than $b_2 = b_{Y2.1} = 0.0866$.

The remainder of this section is devoted to proofs of the basic results (13.3.2) and (13.3.3). Recall that

$$\begin{aligned}\hat{y} &= \hat{Y} - \bar{Y} = b_1x_1 + b_2x_2 : y = \hat{y} + d \\ d &= y - b_1x_1 - b_2x_2\end{aligned}$$

Start with the normal equations:

$$\begin{aligned}b_1 \Sigma x_1^2 + b_2 \Sigma x_1x_2 &= \Sigma x_1y \\ b_1 \Sigma x_1x_2 + b_2 \Sigma x_2^2 &= \Sigma x_2y\end{aligned}$$

These may be rewritten in the form

$$\Sigma x_1(y - b_1x_1 - b_2x_2) = \Sigma x_1d = 0 \quad (13.3.4)$$

$$\Sigma x_2(y - b_1x_1 - b_2x_2) = \Sigma x_2d = 0 \quad (13.3.5)$$

These results show that the deviations d have zero sample correlations with any X -variable. This is not surprising, since d represents the part of Y that is not linearly related either to X_1 or to X_2 .

Multiply (13.3.4) by b_1 and (13.3.5) by b_2 and add. Then

$$\Sigma(b_1x_1 + b_2x_2)d = \Sigma \hat{y}d = 0 \quad (13.3.6)$$

Now

$$\begin{aligned}\Sigma y^2 &= \Sigma (\hat{y} + d)^2 = \Sigma \hat{y}^2 + 2\Sigma \hat{y}d + \Sigma d^2 \\ &= \Sigma \hat{y}^2 + \Sigma d^2\end{aligned}$$

using (13.3.6). This proves the first result (13.3.2). To obtain the second result, we have

$$\begin{aligned}\Sigma \hat{y}^2 &= \Sigma (b_1 x_1 + b_2 x_2)^2 \\ &= b_1^2 \Sigma x_1^2 + 2b_1 b_2 \Sigma x_1 x_2 + b_2^2 \Sigma x_2^2\end{aligned}$$

Reverting to the normal equations, multiply the first one by b_1 , the second by b_2 and add. This gives

$$b_1^2 \Sigma x_1^2 + 2b_1 b_2 \Sigma x_1 x_2 + b_2^2 \Sigma x_2^2 = b_1 \Sigma x_1 y + b_2 \Sigma x_2 y$$

This establishes (13.3.3), the shortcut method of computing the reduction $\Sigma \hat{y}^2$ in *S.S.* due to regression.

EXAMPLE 13.3.1—Here is a set of ten triplets for easy computation.

X_1	X_2	Y	X_1	X_2	Y	
29	2	22	16	1	12	
1	4	26	26	1	13	
5	3	23	15	4	30	
27	1	11	6	2	12	
25	3	25	10	3	26	
			Sums	160	24	200

(i) Calculate the regression, $\hat{Y} = 0.241X_1 + 6.829X_2 - 0.239$

(ii) Predict the value of Y for the fourth member of the sample, ($X_1 = 27$, $X_2 = 1$).
Ans. 13.07.

EXAMPLE 13.3.2—In the preceding example, compute the total *S.S.* of Y and the *S.S.* due to regression. Hence, find the sum of squares of deviations. Ans. 35.0.

EXAMPLE 13.3.3—Show that after allowing for the effects of the other variable, both X_1 and X_2 have a significant relation with Y .

EXAMPLE 13.3.4—Note that when X_1 is fitted alone, the regression coefficient is negative; i.e., Y tends to decrease as X_1 increases. When X_2 is included, the coefficient b_1 becomes significantly positive. From the normal equations the following relation may be proved:

$$b_{Y1 \cdot 2} = b_{Y1} - b_{Y2 \cdot 1} b_{21}$$

where $b_{21} = \Sigma x_1 x_2 / \Sigma x_1^2$ is the regression of X_2 on X_1 . If $b_{Y2 \cdot 1}$ is positive and b_{21} is negative, as in this example, the term $-b_{Y2 \cdot 1} b_{21}$ is positive. If this term is large enough it can change a negative b_{Y1} into a positive $b_{Y1 \cdot 2}$.

13.4—Alternative method of calculation. The inverse matrix. For many purposes, including the construction of confidence intervals for the β 's and the making of comparisons among the b 's, some additional quantities must be computed. If it is known that these will be needed, the calculations given in preceding sections are usually altered slightly, as will be described.

On the left side of the normal equations, the quantities Σx_1^2 , $\Sigma x_1 x_2$, and Σx_2^2 appear. The array

$$\begin{pmatrix} \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_1 x_2 & \Sigma x_2^2 \end{pmatrix}$$

is called a *matrix* with 2 rows and 2 columns—the matrix of sums of squares and products. Mathematicians have defined the *inverse* of this matrix, this being an extension to two dimensions of the concept of the reciprocal of a number. The inverse is also a 2×2 matrix:

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

The elements c_{ij} , called also the *Gauss multipliers*, are found by solving two sets of equations

First Set	Second Set
$c_{11}\Sigma x_1^2 + c_{12}\Sigma x_1 x_2 = 1$	$c_{21}\Sigma x_1^2 + c_{22}\Sigma x_1 x_2 = 0$
$c_{11}\Sigma x_1 x_2 + c_{12}\Sigma x_2^2 = 0$	$c_{21}\Sigma x_1 x_2 + c_{22}\Sigma x_2^2 = 1$

The left side of each set is the same as that of the normal equations. The right sides have 1, 0, or 0, 1, respectively. The first set give c_{11} and c_{12} , the second set c_{21} and c_{22} . It is easy to show that $c_{12} = c_{21}$.

In the 2×2 case the solutions are:

$$c_{11} = \Sigma x_2^2 / D : c_{12} = c_{21} = -\Sigma x_1 x_2 / D : c_{22} = \Sigma x_1^2 / D,$$

where, as before,

$$D = (\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2$$

Note that the numerator of c_{11} is Σx_2^2 , not Σx_1^2 . Note also the negative sign in c_{12} .

In the example, the matrix of sums of squares and products was

$$\begin{pmatrix} 1,752.96 & 1,085.61 \\ 1,085.61 & 3,155.78 \end{pmatrix}$$

with $D = 4,353,400$. This gives

$$c_{11} = 3,155.78 / 4,353,400 = 0.0007249$$

$$c_{12} = -1,085.61 / 4,353,400 = -0.0002494$$

$$c_{22} = 1,752.96 / 4,353,400 = 0.0004027$$

From the c 's, the b 's are obtained as the sums of products of the c 's with the right sides of the normal equations, as follows:

$$\begin{aligned} b_1 &= c_{11}\Sigma x_1 y + c_{12}\Sigma x_2 y \\ &= (0.0007249)(3,231.48) + (-0.0002494)(2,216.44) = 1.7897 \quad (13.4.1) \end{aligned}$$

$$\begin{aligned} b_2 &= c_{21}\Sigma x_1 y + c_{22}\Sigma x_2 y \\ &= (-0.0002494)(3,231.48) + (0.0004027)(2,216.44) = 0.0866 \quad (13.4.2) \end{aligned}$$

The main reason for finding the c 's is that they provide the variances and the covariance of the b 's. The formulas are:

$$V(b_1) = \sigma^2 c_{11} \quad : \quad V(b_2) = \sigma^2 c_{22} \quad : \quad \text{Cov}(b_1 b_2) = \sigma^2 c_{12}$$

where σ^2 is the variance of the residuals of Y from the regression plane.

To summarize, if the c 's are wanted, they are computed first from the normal equations; then the b 's are computed from the c 's as above. The deviations sum of squares and the analysis of variance follow as in section 13.3. Some uses of the c 's are presented in the next section.

EXAMPLE 13.4.1—To prove the relations $b_1 = c_{11}\Sigma x_1 y + c_{12}\Sigma x_2 y$: $b_2 = c_{21}\Sigma x_1 y + c_{22}\Sigma x_2 y$, use these relations to substitute for b_1 and b_2 in terms of the c 's in the left side of the first normal equation. Then show, by the first equation satisfied by the c 's in each set, that this left side equals $\Sigma x_1 y$. Similarly, you can show that the left side of the second normal equation equals $\Sigma x_2 y$. This proves that the b 's computed as above are solutions of the normal equations.

EXAMPLE 13.4.2—Show (i) that b_1 and b_2 have zero correlation only if $\Sigma x_1 x_2 = 0$: (ii) that, in this event, the regression coefficient of Y on X_1 is the same whether X_2 is included in the regression or not. This is the condition that holds for the main effects of each factor in a factorial experiment.

13.5—Standard errors of estimates in multiple regression. In section 13.3 we found that the deviations mean square s^2 was 427.6 with 15 *d.f.*, giving $s = 20.7$. The standard errors of b_1 and b_2 are therefore

$$s_{b_1} = s\sqrt{c_{11}} = 20.7\sqrt{0.0007249} = (20.7)(0.0269) = 0.557 \quad (13.5.1)$$

$$s_{b_2} = s\sqrt{c_{22}} = 20.7\sqrt{0.0004027} = (20.7)(0.0201) = 0.416 \quad (13.5.2)$$

It can be proved that the quantity $(b_1 - \beta_1)/s_{b_1}$ is distributed as t with $(n - k)$ or 15 *d.f.* The null hypothesis $\beta_1 = 0$ can be tested as usual:

$$t_1 = b_1/s_{b_1} = 1.7898/0.557 = 3.21^{**}$$

$$t_2 = b_2/s_{b_2} = 0.0866/0.416 = 0.21$$

These t -tests are identical to the F -tests of the same hypotheses made in table 13.3.2. Note that $(3.21)^2 = 10.30$ and $(0.21)^2 = 0.04$, these being the two values of F found in table 13.3.2.

Evidently in the population of soils that were sampled the fraction of inorganic phosphorus is the better predictor of the plant-available phosphorus. The experiment indicates "that soil organic phosphorus *per se* is not available to plants. Presumably, the organic phosphorus is of appreciable availability to plants only upon mineralization, and in the experiments the rate of mineralization at 20°C. was too low to be of measurable importance."

Confidence limits for any β_i are found as usual. For β_1 , 95% limits are

$$b_1 \pm t_{0.05} s_{b_1} = 1.790 \pm (2.131)(0.557) = 0.60 \text{ and } 2.98$$

Sometimes, comparisons among the b_i are of interest. The standard error of any comparison $\Sigma L_i b_i$ is

$$s\sqrt{(\Sigma L_i)^2 c_{ii} + 2\Sigma L_i L_j c_{ij}} \quad (13.5.3)$$

For example, the standard error of $(b_1 - b_2)$ is

$$s\sqrt{(c_{11} + c_{22} - 2c_{12})} = (20.7)\sqrt{0.0007249 + 0.0004027 - 2(-0.0002494)} \\ = (20.7)\sqrt{(0.0016264)} = 0.835$$

When the regression is constructed for purposes of prediction, we wish to know how accurately \hat{Y} predicts the population mean of Y for specified values of X_1 and X_2 . Call this mean μ . For instance, we might predict the average weight of 11-year-old boys of specified height X_1 and chest girth X_2 . The formula for the estimated standard error of $\hat{Y} = \hat{\mu}$ is

$$s_{\hat{\mu}} = s\sqrt{(1/n + c_{11}x_1^2 + c_{22}x_2^2 + 2c_{12}x_1x_2)} \quad (13.5.4)$$

Example: For the value of Y at the point $X_1 = 4.7$, $X_2 = 24$ (soil sample 5 in table 13.2.1): $x_1 = 4.7 - 11.9 = -7.2$, $x_2 = 24 - 42.1 = -18.1$; so the standard error of the estimate \hat{Y} is

$$(20.7)\sqrt{[1/18 + (0.0007249)(-7.2)^2 + (0.0004027)(-18.1)^2 \\ + 2(-0.0002494)(-7.2)(-18.1)]} = \pm 8.25 \text{ ppm}$$

Alternatively, \hat{Y} may be used to predict the value of Y for an individual new member Y' of the population (that is, one not included in the regression calculations.) In this case,

$$s_t = s\sqrt{1 + \frac{1}{n} + c_{11}x_1^2 + c_{22}x_2^2 + 2c_{12}x_1x_2} \quad (13.5.5)$$

This result is subject to the assumption that the new member comes from the same population as the original data. Unless the predictions satisfy this condition, the standard error should be regarded as tentative. It will be too low if the passage of time or changes in the environment have changed the values of the β 's. If numerous predictions are being made, a direct check on their accuracy should be made whenever possible.

Finally, the standard error of $(Y_i - \hat{Y})$, where Y_i is one of the observations from which the regression was computed, is $\sigma\sqrt{g}$, where

$$g = 1 - \frac{1}{n} - c_{11}x_1^2 - c_{22}x_2^2 - 2c_{12}x_1x_2 \quad (13.5.6)$$

However, if the deviation $(Y_i - \hat{Y})$ has aroused attention because it looks suspiciously large, we cannot apply a t -test of the form $t = (Y_i - \hat{Y})/s\sqrt{g}$, for two reasons. The quantities $(Y_i - \hat{Y})$ and s are not independent, since $(Y_i - \hat{Y})^2$ is a part of the deviations S.S. Secondly, we must allow for the fact that $(Y_i - \hat{Y})$ was picked out because it looks large.

A test can be made as follows. The quantity

$$s'^2 = [\Sigma(Y - \hat{Y})^2 - (Y_i - \hat{Y})^2/g]/(n - k - 1)$$

can be shown to be the mean square of the deviations obtained if the suspect Y_i is omitted when fitting the regression. If Y_i were a randomly chosen observation, the quantity $t' = (Y_i - \hat{Y})/s'\sqrt{g}$ would follow the t -distribution with $(n - k - 1)$ d.f. To make approximate allowance for

the fact that we selected the largest absolute deviation, we regard the deviation as significant at the 5% level if t' is significant at the level $0.05/n$ (This may require reference to detailed tables (3) of t .)

To illustrate, it was noted (section 13.2) that the deviation +58.8 for soil 17 is outstanding. The value of g is found to be 0.80047, while $\Sigma(Y - \hat{Y})^2$ is 6,414 (section 13.3) with 15 $d.f.$ Hence,

$$s'^2 = \frac{1}{14} \left[6,414 - \frac{(58.8)^2}{0.80047} \right] = (6,414 - 4,319)/14 = 150$$

$$t' = (58.8)/\sqrt{(150)(0.80047)} = 5.36 \quad (14 \text{ } d.f.)$$

Since $0.05/18 = 0.0028$, the question is whether a value of 5.36 exceeds the 0.0028 level of t with 14 $d.f.$ Appendix table A 4 shows that the 0.001 level of t is 4.140. The deviation is clearly significant after allowance for the fact that it is the largest. If the regression is recomputed with soil 17 excluded, the main conclusion is not altered. The value of b_1 drops to 1.290 but remains significant, while b_2 becomes -0.111 (non-significant).

EXAMPLE 13.5.1—In the phosphorus data, set 95% confidence limits for β_2 . Ans. -0.79 to 0.97 ppm.

EXAMPLE 13.5.2—For a new soil having $X_1 = 14.6$, $X_2 = 51$, predict the value of Y and give the standard error of your prediction. Ans. $\hat{Y} = 61.86$, s.e. = ± 21.5 ppm, using formula 13.5.5.

EXAMPLE 13.5.3—If Y_i is one of the observations from which the regression was computed, the variance of $Y_i - \hat{Y}$ is (formula 13.5.6),

$$\sigma^2 \left[1 - \frac{1}{n} - c_{11}x_1^2 - c_{22}x_2^2 - 2c_{12}x_1x_2 \right]$$

If this expression is added over all the n sample values, we get

$$\sigma^2 [n - 1 - c_{11}\Sigma x_1^2 - c_{22}\Sigma x_2^2 - 2c_{12}\Sigma x_1x_2]$$

From the equations for the c 's, show that the above equals $\sigma^2(n - 3)$. This is one way of seeing that $\Sigma(Y - \hat{Y})^2$ has $(n - 3)$ $d.f.$

EXAMPLE 13.5.4—With soil 17 omitted, we have

$$\begin{array}{lll} \Sigma X_1 = 188.2 & \Sigma X_2 = 700 & \Sigma Y = 1,295 \\ \Sigma x_1^2 = 1,519.30; & \Sigma x_1x_2 = 835.69, & \Sigma x_2^2 = 2,888.47; \\ \Sigma x_1y = 1,867.39, & \Sigma x_2y = 757.47; & \Sigma y^2 = 4,426.48 \end{array}$$

Solve the normal equations and verify that $b_1 = 1.290$, $b_2 = -0.111$, deviations $SS = 2.101$.

13.6—The interpretation of regression coefficients. In the many areas of research in which controlled experiments are not practicable, multiple regression analyses are extensively used in attempts to disentangle and measure the effects of different X -variables on some response Y . There

are, however, important limitations on what can be learned from this technique in observational studies. While the discussion will be given for a regression on two X -variables, the conclusions apply also when there are more than two. The multiple linear regression model on which the analysis is based is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad (13.6.1)$$

where the residuals ε are assumed to be distributed, independently of the X 's, with zero mean and variance σ^2 . (The assumption of normality of the ε 's is required for tests of significance, but not for the other standard properties of regression estimates.) We assume that the X 's remain fixed in repeated sampling.

In an observational study the investigator looks for some suitable source in which he can measure or record a sample of the triplets (X_1, X_2, Y) . He may try to select the pairs (X_1, X_2) according to some plan, for instance so as to ensure that both X 's vary over a substantial range and that their correlation is not too high, though he is limited in this respect by what the available source can provide.

Difficulty arises because he can never be sure that there are not other X -variables related to Y in the population sampled. These may be variables that he thinks are unimportant, variables that are not feasible to measure or record, or variables unknown to him. Consequently, instead of (13.6.1) the correct regression model is likely to be of the form

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k + \varepsilon'$$

where $X_3 \dots X_k$ represent these additional variables, and k may be fairly large. To keep the algebra simple we replace the additional terms in the model, $\beta_3 X_3 + \dots + \beta_k X_k$, by a single term $\beta_o X_o$, which stands for the joint effect of all the terms omitted from the two-variable model. Thus the correct model is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_o X_o + \varepsilon' \quad (13.6.2)$$

where ε' represents that part of Y that is distributed independently of X_1 , X_2 , and X_o .

The investigator computes the sample regression of Y on X_1 and X_2 as in preceding sections, obtaining the regression coefficients b_1 and b_2 . Under the correct model (13.6.2), it will be proved later that b_1 is an unbiased estimate, not of β_1 , but of

$$\beta_1 + \beta_o b_{o1 \cdot 2} \quad (13.6.3)$$

where $b_{o1 \cdot 2}$ is the sample regression coefficient of X_o on X_1 , after allowing for the effects of X_2 . Clearly, b_1 may be either an overestimate or an underestimate of β_1 . Since the bias in b_1 depends on variables that have not been measured, it is hard to form a judgment about the amount of bias.

For example, an investigator might try to estimate the effects of nitrogen and phosphorus fertilizers on the yield of a common farm

crop by taking a sample of farms. On each field he records the crop yield Y at the most recent harvest and the amounts X_1 , X_2 of N and P per acre applied in that field. If, however, substantial amounts of fertilizer are used mainly by the more competent farmers, the fields on which X_1 and X_2 have high values will, in general, have better soil, more potash fertilizer, superior drainage and tillage, more protection against insect and crop damage, and so on. If $\beta_o X_o$ denotes the combined effect of these variables on yield, X_o will be positively correlated with X_1 and X_2 , so that $b_{o1 \cdot 2}$ will be positive. Further, β_o will be positive if these practices increase yields. Thus the regression coefficients b_1 and b_2 will overestimate the increase in yield caused by additional amounts of N and P . This type of overestimation is likely to occur whenever the beneficial effects of an innovation in some process are being estimated by regression analysis, if the more capable operators are the ones who try out the innovation.

When the purpose is to find a regression formula that predicts Y accurately rather than to interpret individual regression coefficients, the bias in b_1 may actually be advantageous. Insofar as the unknown variables in X_o are good predictors of Y and are stably related to X_1 , the regression value of b_1 is in effect trying to improve the prediction by capitalizing on these relationships. This can be seen from an artificial example (in which X_2 is omitted for simplicity). Suppose that the correct model is $Y = 1 + 3X_o$. This implies that in the correct model (i) X_1 is useless as a predictor, since $\beta_1 = 0$, (ii) if X_o could be measured, it would give perfect predictions, since the model has no residual term ε' . In the data (table 13.6.1), we have constructed an X_1 that is highly correlated with X_o . You may check that the prediction equation based on the regression of Y on X_1 ,

$$\hat{Y}_1 = 2.5 + 3.5X_1$$

gives good, although not perfect, predictions. Since $\beta_o = 3$, $b_{o1} = 7/6$, $b_1 = 7/2$, the relation $b_1 = \beta_o b_{o1}$ is also verified.

TABLE 13.6 1
ARTIFICIAL EXAMPLE TO ILLUSTRATE PREDICTION FROM AN INCOMPLETE REGRESSION MODEL

Observation	X_o	$Y = 1 + 3X_o$	X_1	\hat{Y}_1	$Y - \hat{Y}_1$
1	1	4	0	2.5	+1.5
2	2	7	2	9.5	-2.5
3	4	13	3	13.0	0.0
4	6	19	5	20.0	-1.0
5	7	22	5	20.0	+2.0
Sum	20	65	15	65.0	0.0
Mean	4	13	3	13.0	0.0

$$\Sigma x_1^2 = 18, \quad \Sigma x_1 y = 63, \quad \Sigma x_o x_1 = 21, \quad b_1 = \frac{63}{18} = 3.5, \quad b_{o1} = \frac{21}{18} = \frac{7}{6}$$

To return to studies in which the sizes of the regression coefficients are of primary interest, a useful precaution is to include in the regression any X -variable that seems likely to have a material effect on Y , even though this variable is not of direct interest. Note from formula 13.6.3 that no contribution to the bias in b_1 comes from β_2 , since X_2 was included in the regression. Another strategy is to find, if possible, a source population in which X -variables not of direct interest have only narrow ranges of variation. The effect is to decrease $b_{01.2}$ (see example 13.6.1) and hence lessen the bias in b_1 . It also helps if the study is repeated in diverse populations that are subject to different X_0 variables. The finding of stable values for b_1 and b_2 gives reassurance that the biases are not major.

In many problems the variables X_1 and X_2 are thought to have causal effects on Y . We would like to learn by how much Y will be increased (if beneficial) or decreased (if harmful) by a given change ΔX_1 in X_1 . The estimate of this amount suggested by the multiple regression equation is $b_1 \Delta X_1$. As we have seen, this quantity is actually an estimate of $(\beta_1 + \beta_0 b_{01.2}) \Delta X_1$. Further, while we may be able to impose a change of amount ΔX_1 in X_1 we may be unable to control other consequences of this change. These consequences may include changes ΔX_2 in X_2 and ΔX_0 in X_0 . Thus the real effect of a change ΔX_1 may be, from model 13.6.2,

$$\beta_1 \Delta X_1 + \beta_2 \Delta X_2 + \beta_0 \Delta X_0, \quad (13.6.4)$$

whereas our estimate of this amount, which assumes that ΔX_1 can be changed without producing a change in X_2 and ignores the unknown variables, approximates $(\beta_1 + \beta_0 b_{01.2}) \Delta X_1$. If enough is known about the situation, a more realistic mathematical model can be constructed, perhaps involving a system of equations or path analysis (26, 27). In this way a better estimate of 13.6.4 might be made, but estimates of this type are always subject to hazard. As Box (4) has remarked, in an excellent discussion of this problem in industrial work, "To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it)."

To sum up, when it is important to find some way of increasing or decreasing Y , multiple regression analyses provide indications as to which X -variables might be changed to accomplish this end. Our advance estimates of the effects of such changes on Y , however, may be wrong by substantial amounts. If these changes are to be imposed, we should plan, whenever feasible, a direct study of the effects of the changes on Y so that false starts can be corrected quickly.

In controlled experiments these difficulties can be largely overcome. The investigator is able to impose the changes (treatments) whose effects he wishes to measure and to obtain direct measurements of their effects. The extraneous and unknown variables represented by X_0 are present just as in observational studies. But the device of randomization (5, 6) makes λ_0 in effect independent of X_1 and X_2 in the probability sense. Thus X_0

acts like the residual term ε in the standard regression model and the assumptions of this model are more nearly satisfied. If the effects of X_o are large, the Deviations mean square, which is used as the estimate of error, will be large, and the experiment may be too imprecise to be useful. A large error variance should lead the investigator to study the uncontrolled sources of variation in order to find a way of doing more accurate experimentation.

We conclude this section with a proof of the result (equation 13.6.3): namely, that if a regression of Y on X_1 and X_2 is computed under the model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_o X_o + \varepsilon' \quad E(\varepsilon') = 0,$$

then

$$E(b_1) = \beta_1 + \beta_o b_{o1.2} \quad (13.6.3)$$

The result is important in showing that a regression coefficient is free from any bias due to other X 's like X_2 that are included in the fitted regression, but is subject to bias from X 's that were omitted. Since it is convenient to work with deviations from the sample means, note that from the model, we have

$$y = Y - \bar{Y} = \beta_1 x_1 + \beta_2 x_2 + \beta_o x_o + \varepsilon' - \bar{\varepsilon}' \quad (13.6.5)$$

Now,

$$b_1 = c_{11} \Sigma x_1 y + c_{12} \Sigma x_2 y$$

Substitute for y from 13.6.5,

$$\begin{aligned} b_1 = & c_{11} \Sigma x_1 (\beta_1 x_1 + \beta_2 x_2 + \beta_o x_o + \varepsilon' - \bar{\varepsilon}') \\ & + c_{12} \Sigma x_2 (\beta_1 x_1 + \beta_2 x_2 + \beta_o x_o + \varepsilon' - \bar{\varepsilon}') \end{aligned}$$

When we average over repeated samples, all terms in ε' , like $c_{11} \Sigma x_1 \varepsilon'$, vanish because ε' has mean zero independently of x_1 , x_2 , and x_o . Collect terms in β_1 , β_2 , and β_o .

$$\begin{aligned} E(b_1) = & \beta_1 (c_{11} \Sigma x_1^2 + c_{12} \Sigma x_1 x_2) + \beta_2 (c_{11} \Sigma x_1 x_2 + c_{12} \Sigma x_2^2) \\ & + \beta_o (c_{11} \Sigma x_1 x_o + c_{12} \Sigma x_2 x_o) \end{aligned}$$

From the first set of equations satisfied by c_{11} and c_{12} (section 13.4), the coefficient of β_1 is 1 and that of β_2 is 0.

What about the coefficient of β_o ? Notice that it resembles

$$c_{11} \Sigma x_1 y + c_{12} \Sigma x_2 y = b_1,$$

except that x_o has replaced y . Hence, the coefficient of β_o is the regression coefficient $b_{o1.2}$ of X_o on X_1 that would be obtained by computing the sample regression of X_o on X_1 and X_2 . This completes the proof.

EXAMPLE 13.6.1.—This illustrates the result that when there are omitted variables denoted by X_o , the bias that they create in b_1 depends both on the size β_o of their effect on Y and on the extent to which X_o varies. Let $Y = X_1 + X_o$, so that $\beta_1 = \beta_o = 1$. In sample 1,

X_1 and X_0 have the same distribution. Verify that $b_1 = 2$. In sample 2, X_1 and X_0 still have a perfect correlation but the variance of X_0 is greatly reduced. Verify that b_1 is now 1.33, giving a much smaller bias. Of course, steps that reduce the correlation between X_1 and X_0 are also helpful.

Sample 1				Sample 2			
X_1	X_0	Y		X_1	X_0	Y	
-6	-6	-12		-6	-2	-8	
-3	-3	-6		-3	-1	-4	
0	0	0		0	0	0	
0	0	0		0	0	0	
9	9	18		9	3	12	
Sum	0	0	0	Sum	0	0	0
$\Sigma x_1^2 = 126, \Sigma x_1 y = 252$				$\Sigma x_1^2 = 126, \Sigma x_1 y = 168$			

13.7—Relative importance of different X-variables. In a multiple-regression analysis the question may be asked: Which X variables are most important in determining Y ? Usually, no unique or fully satisfactory answer can be given, but several approaches have been tried. Consider first the situation in which the objective is to predict Y or to “explain” the variation in Y . The problem would be fairly straightforward if the X -variables were independent. From the model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

we have, in the population,

$$\sigma_Y^2 = \beta_1^2 \sigma_1^2 + \beta_2^2 \sigma_2^2 + \dots + \beta_k^2 \sigma_k^2 + \sigma^2$$

where σ_i^2 denotes the variance of X_i . The quantity $\beta_i^2 \sigma_i^2 / \sigma_Y^2$ measures the fraction of the variance of Y attributable to its linear regression on X_i . This fraction can be reasonably regarded as a measure of the relative importance of X_i . With a random sample from this population, the quantities $b_i^2 \Sigma x_i^2 / \Sigma y^2$ are sample estimates of these fractions. (In small samples a correction for bias might be advisable since $b_i^2 \Sigma x_i^2 / \Sigma y^2$ is not an unbiased estimate of $\beta_i^2 \sigma_i^2 / \sigma_Y^2$.)

The square roots of these quantities, $b_i \sqrt{(\Sigma x_i^2 / \Sigma y^2)}$, called the *standard partial regression coefficients*, have sometimes been used as measures of relative importance, the X 's being ranked in order of the sizes of these coefficients (ignoring sign). The quantity $\sqrt{(\Sigma x_i^2 / \Sigma y^2)}$ is regarded as a correction for scale. The coefficient estimates $\beta_i \sigma_i / \sigma_Y$, the change in Y , as a fraction of σ_Y , produced by one S.D. change in X_i .

In practice, correlations between the X 's make the answer more difficult. In many applications, X_1 and X_2 are positively correlated with each other and with Y . For instance, X_1 and X_2 may be examination scores that predict a student's ability to do well in a course, and Y his final score in that course. To illustrate this case, table 13.7.1 shows the normal

equations, the b 's and the analysis of variance. As the example is constructed, X_1 is a slightly better predictor than X_2 , the two together accounting for about 70% of the variation in Y (reduction due to regression 26.53 out of a total $S.S.$ of 38.00).

As is typical in such applications, each variable's contribution to Σy^2 is much greater when the variable is used alone than when it follows the other variable. For X_1 the two sums of squares are 22.50 and 9.63, respectively, while for X_2 they are 16.90 and 4.03. If the sums of squares when X_1 and X_2 appear alone are taken to measure the contributions of X_1 and X_2 to the variation in Y , the two contributions add to 39.40, which is more than Σy^2 (38.00). On the other hand the sums of squares 9.63 and 4.03 greatly underestimate the joint contribution of X_1 and X_2 . Neither method of measuring the relative contribution is satisfactory.

TABLE 13.7.1
A COMMON SITUATION IN TWO-VARIABLE REGRESSION. ARTIFICIAL DATA

Normal equations:

$$\begin{aligned} 10b_1 + 5b_2 &= 15 \\ 5b_1 + 10b_2 &= 13 \\ c_{11} = c_{22} = 2/15 : c_{12} &= -1/15 : b_1 = 17/15 : b_2 = 11/15 \end{aligned}$$

Source of Variation	Degrees of Freedom	Sum of Squares
Total	52	38.00
Regression on X_1 alone	1	$(\Sigma x_1 y)^2 / \Sigma x_1^2 = 15^2 / 10 = 22.50$
Regression on X_2 after X_1	1	$b_2^2 / c_{22} = 11^2 / 30 = 4.03$
Regression on X_2 alone	1	$(\Sigma x_2 y)^2 / \Sigma x_2^2 = 13^2 / 10 = 16.90$
Regression on X_1 after X_2	1	$b_1^2 / c_{11} = 17^2 / 30 = 9.63$
Deviation	50	11.47

Sometimes the investigator's question is: Is X_1 when used alone a better predictor of Y than X_2 when used alone? In this case, comparison of the numbers 22.50 and 16.90 is appropriate. An answer to the question has been given by Hotelling (7) for two X -variables and extended by Williams (8) to more than two.

In other applications there may be a rational way of deciding the order in which the X 's should be brought into the regression, so that their contributions to Σy^2 add up to the correct combined contribution. In his studies of the variation in the yields of wheat grown continuously on the same plots for many years at Rothamsted, Fisher (2) postulated the sources of variation in the following order: (1) A steady increase or decrease in level of yield, measured by a linear regression on time; (2) other slow changes in yields through time, represented by a polynomial in time with terms in T^2 , T^3 , T^4 , T^5 ; (3) the effect of total annual rainfall on the deviations of yields from the temporal trend; (4) the effect of the dis-

tribution of rainfall throughout the growing season on the deviations from the preceding regression.

Finally, if the purpose is to learn how to change Y in some population by changing some X -variable, the investigator might estimate the sizes ΔX_1 , ΔX_2 , etc., of the changes that he can impose on X_1 and X_2 in this population by a given expenditure of resources. He might then rate the variables in the order of the sizes of $b_i \Delta X_i$, in absolute terms, these being the estimated amounts of change that will be produced in Y . As we have seen in the preceding section, this approach has numerous pitfalls.

13.8—Partial and multiple correlation. In a sample of 18-year-old college freshmen, the variables measured might be height, weight, blood pressure, basal metabolism, economic status, aptitude, etc. One purpose might be to examine whether aptitude (Y) was linearly related to the physiological measurements. If so, the regression methods of the preceding sections would apply. But the objective might be to study the correlations among such variables as height, weight, blood pressure, basal metabolism, etc., among which no variables can be specified as independent or dependent. In that case, *partial correlation* methods are appropriate.

You may recall that the ordinary correlation coefficient was closely related to the bivariate normal distribution. With more than two variables, an extension of this distribution called the *multivariate normal distribution* (9) forms the basic model in correlation studies. A property of the multivariate normal model is that any variable has a linear regression on the other variables (or on any subset of the other variables), with deviations that are normally distributed. Thus, the assumptions made in multivariate regression studies hold for a multivariate normal population.

If there are three variables, there are three simple correlations among them, ρ_{12} , ρ_{13} , ρ_{23} . The *partial correlation coefficient*, $\rho_{12 \cdot 3}$, is the correlation between variables 1 and 2 in a cross section of individuals *all having the same value of variable 3*; the third variable is held constant so that only 1 and 2 are involved in the correlation. In the multivariate normal model, $\rho_{12 \cdot 3}$ is the same for every value of variable 3.

A sample estimate $r_{12 \cdot 3}$ of $\rho_{12 \cdot 3}$ can be obtained by calculating the deviations d_{13} of variable 1 from its sample regression on variable 3. Similarly, find d_{23} . Then $r_{12 \cdot 3}$ is the simple correlation coefficient between d_{13} and d_{23} . The idea is to measure that part of the correlation between variables 1 and 2 that is not simply a reflection of their relations with variable 3. It may be shown that $r_{12 \cdot 3}$ satisfies the following formula:

$$r_{12 \cdot 3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Table A 11 is used to test the significance of $r_{12 \cdot 3}$. Enter it with $(n - 3)$ degrees of freedom, instead of $(n - 2)$ as for a simple correlation coefficient

In Iowa and Nebraska, a random sample of 142 older women was drawn for a study of nutritional status (12). Three of the variables were Age, Blood pressure, and the Cholesterol concentration in the blood. The three simple correlations were

$$r_{AB} = 0.3332, r_{AC} = 0.5029, r_{BC} = 0.2495$$

Since high blood pressure might be associated with above-average amounts of cholesterol in the walls of blood vessels, it is interesting to examine r_{BC} . But it is evident that both B and C increase with age. Are they correlated merely because of their common association with age or is there a real relation at every age? The effect of age is eliminated by calculating

$$r_{BC \cdot A} = \frac{0.2495 - (0.3332)(0.5029)}{\sqrt{(1 - 0.3332^2)(1 - 0.5029^2)}} = 0.1233$$

With $f = 142 - 3 = 139$, this correlation is not significant. It may be that within the several age groups blood pressure and blood cholesterol are uncorrelated. At least, the sample is not large enough to detect the correlation if it is present.

As another illustration, consider the consumption of protein and fat among the 54 older women who came from Iowa. The simple correlations were

$$r_{AP} = -0.4865, r_{AF} = -0.5296, r_{PF} = 0.5784$$

The third correlation shows that protein and fat occur together in all diets while the first two correlations indicate the decreasing quantities of both as age advances; both P and F depend on A . How closely do they depend on each other at any one age?

$$r_{PF \cdot A} = \frac{0.5784 - (-0.4865)(-0.5296)}{\sqrt{(1 - 0.4865^2)(1 - 0.5296^2)}} = 0.4328$$

Part of the relationship depends on age but part of it is inherent in the ordinary composition of foods eaten.

To get a clearer notion of the way in which $r_{PF \cdot A}$ is independent of age, consider the six women near 70 years of age. Their protein and fat intakes were

P :	56,	47,	33,	39,	42,	38	
F :	56,	83,	49,	52,	65,	52	$r_{PF} = 0.4194$

The correlation is close to the average, $r_{PF \cdot A} = 0.4328$. Similar correlations would be found at other ages.

With four variables the partial correlation coefficient between variables 1 and 2 can be computed after eliminating the effects of the other variables, 3 and 4. The formula is

$$r_{12 \cdot 34} = \frac{r_{12 \cdot 4} - r_{13 \cdot 4}r_{23 \cdot 4}}{\sqrt{(1 - r_{13 \cdot 4}^2)(1 - r_{23 \cdot 4}^2)}}$$

or, alternatively,

$$r_{12 \cdot 34} = \frac{r_{12 \cdot 3} - r_{14 \cdot 3}r_{24 \cdot 3}}{\sqrt{(1 - r_{14 \cdot 3}^2)(1 - r_{24 \cdot 3}^2)}},$$

the two formulas being identical.

To test this quantity in table A 11, use $(n - 4)$ degrees of freedom.

As we have stated, partial correlation does not involve the notion of independent and dependent variables: it is a measure of interdependence. On the other hand, the *multiple correlation coefficient* applies to the situation in which one variable, say Y , has been singled out to examine its joint relation with the other variables. In the population, the multiple correlation coefficient between Y and X_1, X_2, \dots, X_k is defined as the simple correlation coefficient between Y and its linear regression, $\beta_1 X_1 + \dots + \beta_k X_k$, on $X_1 \dots X_k$. Since it is hard to attach a useful meaning to the sign of this correlation, most applications deal with its square. The sample estimate R of a multiple correlation coefficient is, as would be expected, the simple correlation between y and $\hat{y} = b_1 x_1 + \dots + b_k x_k$. This gives

$$R^2 = (\Sigma y\hat{y})^2 / (\Sigma y^2)(\Sigma \hat{y}^2)$$

In formula 13.3.6 (p. 388) it was shown that $\Sigma d\hat{y} = 0$, where $d = y - \hat{y}$. It follows that $\Sigma y\hat{y} = \Sigma \hat{y}^2$. Hence,

$$R^2 = \Sigma \hat{y}^2 / \Sigma y^2 \quad : \quad 1 - R^2 = \Sigma d^2 / \Sigma y^2$$

Thus, in the analysis of variance of a multiple regression, R^2 is the fraction of the sum of squares of deviations of Y from its mean that is attributable to the regression, while $(1 - R^2)$ is the fraction not associated with the regression. This result is a natural extension of the corresponding result (section 7.3) for a simple correlation coefficient. The test of the null hypothesis that the multiple correlation in the population is zero is identical to the F -test of the null hypothesis that $\beta_1 = \beta_2 = \dots = \beta_k = 0$. The relation is

$$F = (n - k - 1)R^2 / k(1 - R^2), \text{ with } k \text{ and } (n - k - 1) \text{ d.f.}$$

EXAMPLE 13.8.1—Brunson and Willier (13) examined the correlations among ear circumference E , cob circumference C , and number of rows of kernels K calculated from measurements of 900 ears of corn:

$$r_{EC} = 0.799, \quad r_{EK} = 0.570, \quad r_{CK} = 0.507$$

Among the ears having the same kernel number, what is the correlation between E and C ?
Ans $r_{EC \cdot K} = 0.720$.

EXAMPLE 13.8.2—Among ears of corn having the same circumference, is there any correlation between C and K ? Ans. $r_{CK \cdot E} = 0.105$.

EXAMPLE 13.8.3—In a random sample of 54 Iowa women (12), the intake of two

nutrients was determined together with age and the concentration of cholesterol in the blood. If P symbolizes protein, F fat, A age, and C cholesterol, the correlations are as follows:

	A	P	F
P	-0.4865		
F	-0.5296	0.5784	
C	0.4737	-0.4249	-0.3135

What is the correlation between age and cholesterol independent of the intake of protein and fat? Ans.

$$r_{AC \cdot PF} = \frac{r_{AC \cdot F} - r_{AP \cdot F} r_{CP \cdot F}}{\sqrt{(1 - r_{AP \cdot F}^2)(1 - r_{CP \cdot F}^2)}} = \frac{0.3820 - (-0.2604)(-0.3145)}{\sqrt{(1 - 0.2604^2)(1 - 0.3145^2)}} = 0.3274$$

EXAMPLE 13.8.4—Show that the sample estimate of the fraction of the variance of Y that is attributable to its linear regression on $X_1 \dots X_k$ is

$$1 - \frac{(1 - R^2)(n - 1)}{(n - k - 1)}$$

13.9—Three or more independent variables. Computations. The formulas already described for two X -variables extend naturally to three or more X -variables. The computations inevitably become lengthier: they are ideally suited to an electronic computer. We shall describe one of the standard methods for a desk calculating machine—the *Abbreviated Doolittle method* (10)—except that for clarity more steps are given than an experienced operator needs. For more extensive discussion of computing methods, see (11).

With three independent variables, the normal equations are:

$$\begin{aligned} b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 + b_3 \Sigma x_1 x_3 &= \Sigma x_1 y \\ b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 + b_3 \Sigma x_2 x_3 &= \Sigma x_2 y \\ b_1 \Sigma x_1 x_3 + b_2 \Sigma x_2 x_3 + b_3 \Sigma x_3^2 &= \Sigma x_3 y \end{aligned}$$

If the c 's are needed, as in most applications, the right sides become 1, 0, 0 for c_{11} , c_{12} , c_{13} ; 0, 1, 0 for c_{21} , c_{22} , c_{23} ; and 0, 0, 1 for c_{31} , c_{32} , c_{33} .

Since the same calculating routine can be used for b 's and c 's, only the right sides being different, we denote the unknowns by z_1 , z_2 , z_3 , and let $s_{ij} = \Sigma x_i x_j$. The equations to be solved are:

$$\begin{aligned} (1) \quad & s_{11}z_1 + s_{12}z_2 + s_{13}z_3 = \\ (2) \quad & s_{12}z_1 + s_{22}z_2 + s_{23}z_3 = \\ (3) \quad & s_{13}z_1 + s_{23}z_2 + s_{33}z_3 = \end{aligned}$$

The right side is not specified, since it depends on whether the b 's or c 's are being computed.

The Doolittle method eliminates z_1 , then z_1 and z_2 , solving for z_3 . Intermediate steps provide convenient equations for finding z_2 from z_3 , and finally z_1 from z_2 and z_3 . The computing routine can be carried out

without any thought as to why it works. The explanation is given in this section.

The first step, line (4), is to recopy line (1).

$$(4) \quad s_{11}z_1 + s_{12}z_2 + s_{13}z_3 =$$

Now divide through by s_{11} . It is quicker to find the reciprocal, $1/s_{11}$, and multiply through by $1/s_{11}$. This gives

$$(5) \quad z_1 + \left[\frac{s_{12}}{s_{11}} \right] z_2 + \left\{ \frac{s_{13}}{s_{11}} \right\} z_3 =$$

The coefficients of z_2 and z_3 have been bracketed, since they play a key role. Multiply (4) by s_{12}/s_{11} , obtaining

$$(6) \quad s_{12}z_1 + \frac{s_{12}^2}{s_{11}} z_2 + \frac{s_{12}s_{13}}{s_{11}} z_3 =$$

In steps (5) and (6) and in all subsequent steps, the right side of the equation is always multiplied by the same factor as the left side. Now subtract (6) from (2) to get rid of z_1 .

$$(7) \quad \left(s_{22} - \frac{s_{12}^2}{s_{11}} \right) z_2 + \left(s_{23} - \frac{s_{12}s_{13}}{s_{11}} \right) z_3 =$$

The next operations resemble those in lines (4) to (6). Find the reciprocal of $(s_{22} - s_{12}^2/s_{11})$ and multiply (7) by this reciprocal.

$$(8) \quad z_2 + \left\{ \frac{s_{23} - s_{12}s_{13}/s_{11}}{s_{22} - s_{12}^2/s_{11}} \right\} z_3 =$$

The coefficient of z_3 in (8) receives a curly bracket, like that of z_3 in (5). Reverting to (4) and (5), multiply (4) by the bracketed s_{13}/s_{11} in (5).

$$(9) \quad s_{13}z_1 + \frac{s_{12}s_{13}}{s_{11}} z_2 + \frac{s_{13}^2}{s_{11}} z_3 =$$

Similarly, multiply (7) by the bracketed coefficient of z_3 in (8).

$$(10) \quad \left(s_{23} - \frac{s_{12}s_{13}}{s_{11}} \right) z_2 + \frac{(s_{23} - s_{12}s_{13}/s_{11})^2}{s_{22} - s_{12}^2/s_{11}} z_3 =$$

Now take (3) - (9) - (10). Note that the coefficients of z_1 and z_2 both disappear, leaving an equation (11) with only z_3 on the left. Solve this for z_3 . (If there are four X -variables, continue through another cycle of these operations, ending with an equation in which z_4 alone appears.)

Having z_3 , find z_2 from (8), and finally z_1 from (5). With familiarity, the operator will find that lines (6), (9), and (10) need not be written down when he is using a modern desk machine.

The next two sections give numerical examples of the calculation of the b 's and c 's. The numbering of the lines and all computing instructions in these examples are exactly as in this section.

13.10—Numerical example. Computing the b 's. In table 13.10.1 an additional independent variable X_3 is taken from the original data in the plant-available phosphorus investigation. Like X_2 , the variable X_3 measures organic phosphorus, but of a different type. As before, Y is the estimated plant-available phosphorus in corn grown at Soil Temperature 20°C. The data for Soil Temperature 35°C. are considered later.

TABLE 13.10.1
PHOSPHORUS FRACTIONS IN VARIOUS CALCAREOUS SOILS, AND ESTIMATED PLANT-AVAILABLE
PHOSPHORUS AT TWO SOIL TEMPERATURES

Soil Sample No.	Phosphorus Fractions in Soil, ppm*			Estimated Plant-available Phosphorous in Soil, ppm	
	X_1	X_2	X_3	Soil Temp. 20° C. Y	Soil Temp. 35° C. Y'
1	0.4	53	158	64	93
2	0.4	23	163	60	73
3	3.1	19	37	71	38
4	0.6	34	157	61	109
5	4.7	24	59	54	54
6	1.7	65	123	77	107
7	9.4	44	46	81	99
8	10.1	31	117	93	94
9	11.6	29	173	93	66
10	12.6	58	112	51	126
11	10.9	37	111	76	75
12	23.1	46	114	96	108
13	23.1	50	134	77	90
14	21.6	44	73	93	72
15	23.1	56	168	95	90
16	1.9	36	143	54	82
17	26.8	58	202	168	128
18	29.9	51	124	99	120

* X_1 = inorganic phosphorus by Bray and Kurtz method

X_2 = organic phosphorus soluble in K_2CO_3 and hydrolyzed by hypobromite

X_3 = organic phosphorus soluble in K_2CO_3 and not hydrolyzed by hypobromite

In general, regression problems in which the b 's but not the c 's are wanted are encountered only when the investigator is certain that all the X 's must be present in the regression equation and does not want to test individual b_i or compute confidence limits for any β_i . The present example is a borderline case. A primary objective was to determine whether there exists an independent effect of soil organic phosphorus on the phosphorus nutrition of plants. That is, the investigators wished to

know if X_2 and X_3 are related to Y after allowing for the relation between Y and X_1 (soil inorganic phosphorus). As a first step, we can work out the regression of Y on all three variables, obtaining the reduction in sum of squares of Y . The reduction due to a regression on X_1 alone is $(\sum x_{1y})^2 / \sum x_1^2$. By subtraction, the additional reduction due to a regression on X_2 and X_3 is obtained. It can be tested against the Deviations mean square by an F -test. If F is near 1, this probably settles the issue and the c 's are not needed. But if F is close to its significance level, we will want to examine b_2 and b_3 individually, since one type of inorganic phosphorus might show an independent relation with Y but not the other.

TABLE 13.10.2
SOLUTION OF THREE NORMAL EQUATIONS. ABBREVIATED DOOLITTLE METHOD

Line	Reciprocal	Instructions	X_1	X_2	X_3	Y
(1)			1,752.96	1,085.61	1,200.00	3,231.48
(2)			1,085.61	3,155.78	3,364.00	2,216.44
(3)			1,200.00	3,364.00	35,572.00	7,593.00
(4)		copy (1)	1,752.96	1,085.61	1,200.00	3,231.48
(5)	.03570464	(4) \times .03570464	1	[.61930]	{.68456}	1.84344
(6)		(4) \times [.61930]		672.32	743.16	2,001.26
(7)		(2) - (6)		2,483.46	2,620.84	215.18
(8)	.03402664	(7) \times .03402664		1	{1.05532}	.08665
(9)		(4) \times {.68456}			821.47	2,212.14
(10)		(7) \times {1.05532}			2,765.82	227.08
(11)		(3) - (9) - (10)			31,984.71	5,153.78
		- by 31,984.71 Line (8) Line (5)		$b_2 = 0.8665 - (1.05532)b_3 = -0.08339$ $b_1 = 1.84344 - (.61930)b_2 - (.68456)b_3$ $b_1 = 1.78478$	$b_3 = 0.16113$	
Reduction in $SS = \sum b_i(\sum x_{iy}) = (1.78478)(3,231.48) + + (0.16113)(7,593.00)$ = 6,806						

The normal equations and computation of the b 's are in table 13.10.2. Before starting, consider whether some coding of the normal equations is advisable. If the sizes of the $\sum x_i^2$ differ greatly, it is more difficult to keep track of the decimal places. Division or multiplication of some X 's by a power of 10 will help. If X_i is divided by 10^p , $\sum x_i^2$ is divided by 10^{2p} and $\sum x_i x_j$ or $\sum x_i y$ by 10^p . Note that b_i is multiplied by 10^p and therefore must be divided by 10^p in a final decoding. For practice, see example 13.10.6. In this example no coding seems necessary.

It is hoped that the calculations can be easily followed from the column of Instructions. In the equations like (5) in which the coefficient of the leading b_i is 1, we carried five decimal places, usually enough with

three or four X -variables. Don't forget that the b 's are found in reverse order: b_3 , then b_2 , then b_1 . Since mistakes in calculation are hard to avoid, always substitute the b 's in the original equations as a check, apart from rounding errors. At the end, the reduction in sum of squares of Y is computed.

Table 13.10.3 gives the analysis of variance and the combined test of X_2 and X_3 . Since $F = 1.06$, it seems clear that neither form of inorganic phosphorus is related to Y in these data.

TABLE 13.10.3
ANALYSIS OF VARIANCE AND TEST OF X_2, X_3

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Total	17	12,390		
Regression on X_1, X_2, X_3	3	6,806		
Regression on X_1	1	5,957*		
Regression on X_2, X_3 after X_1	2	849	424	1.06
Deviations	14	5,584	399	

$$* (\Sigma x_1 y)^2 / \Sigma x_1^2 = (3,231.48)^2 / 1,752.96$$

Some general features of multiple regression may now be observed:

1. As noted before, the regression coefficients change with each new grouping of the X . With X_2 alone, $b_{Y2} = 2,216.44/3,155.78 = 0.7023$. Adding X_1 , $b_{Y2.1} = 0.0866$. With three of the X , $b_{Y2.13} = -0.0834$. In any one multiple regression, the coefficients are intercorrelated; either increasing or decreasing the number of X 's changes all the b 's.

2. The value of $\Sigma \hat{y}^2$ never decreases with the addition of new X ; ordinarily it increases. Take X_1 alone; $\Sigma \hat{y}_1^2 = (3,231.48)^2 / 1,752.96 = 5,957$. X_1 and X_2 make $\Sigma \hat{y}_{12}^2 = 5,976$. For all three, $\Sigma \hat{y}_{123}^2 = 6,806$. The increase may be small and nonsignificant, but it estimates the contribution of the added X .

3. For checking calculations it is worth noting that $\Sigma \hat{y}^2$ cannot be greater than Σy^2 ; nearly always it is less. Only if the X predict Y perfectly can $\Sigma \hat{y}^2 = \Sigma y^2$. In that limiting case, $\Sigma d^2 = 0$.

4. High correlation between two of the X can upset calculations. If r_{ij} is above 0.95, even 6 or 8 significant digits may not be sufficient to control rounding errors. Consider eliminating one of the two X 's.

5. If $\Sigma \hat{y}^2$ is only a small fraction of Σy^2 , that is, if R^2 is small, remember that most of the variation in Y is unexplained. It may be random variation or it may be due to other independent variables not considered in the regression. If these other variables were found and brought in, the relations among the X 's already included might change completely.

EXAMPLE 13.10.1—Compute the regression of plant-available phosphorus on the 3 fractions. Ans. $\hat{Y} = 1.7848X_1 - 0.0834X_2 + 0.1611X_3 + 43.67$.

EXAMPLE 13.10.2—Estimate the plant-available phosphorus in soil sample 17 and compare it with the observed value. Ans. 119 ppm., $Y - \hat{Y} = 49$ ppm.

EXAMPLE 13.10.3—The experimenter might have information which would lead him to retain X_3 along with X_1 in his predicting equation, dropping X_2 . Calculate the new regression. Ans. $\hat{Y} = 1.737X_1 + 0.155X_3 + 41.5$.

EXAMPLE 13.10.4—Calculate the sum of squares due to X_2 after X_1 and X_3 . Ans. 16.

EXAMPLE 13.10.5—Calculate $R^2 = \Sigma \hat{y}^2 / \Sigma y^2$ with X_1 alone, with X_1 and X_2 , and with X_1, X_2, X_3 . Ans. $R_{Y \cdot 1}^2 = 0.4808$, $R_{Y \cdot 12}^2 = 0.4823$, $R_{Y \cdot 123}^2 = 0.5493$. Notice that R^2 never decreases with the addition of a new X ; ordinarily it increases. Associate this with the corresponding theorem about $\Sigma \hat{y}^2$.

EXAMPLE 13.10.6—In a multiple regression the original normal equations were as follows:

X_1	X_2	X_3	Y
1.28	17.20	85.20	2.84
17.20	2,430.00	7,160.00	183.00
85.20	7,160.00	67,200.00	8,800.00

It was decided to divide X_2 by 10 and X_3 by 100 before starting the solution. What happens to ΣX_1X_3 , ΣX_2Y , ΣX_2X_3 , ΣY^2 , ΣX_2^2 , ΣX_3Y ? Ans. They become 0.852, 18.30, 7.16, 6.72, 24.30, 88.00.

EXAMPLE 13.10.7—In studies of the fertilization of red clover by honey bees (28), it was desired to learn the effects of various lengths of the insects' probosces. The measurement is difficult, so a pilot experiment was performed to determine a more convenient one that might be highly correlated with proboscis length. Three measurements were tried on 44 bees with the results indicated:

$n = 44$	Dry Weight, X_1 (mg.)	Length of Wing, X_2 (mm.)	Width of Wing, X_3 (mm.)	Length of Proboscis, Y (mm.)
Mean	13.10	9.61	3.28	6.59
Sum of Squares and Products				
	X_1	X_2	X_3	Y
X_1	16 6840	1.9279	0.8240	1.5057
X_2		0.9924	0.3351	0.5989
X_3			0.2248	0.1848
Y				0 6831

Coding is scarcely necessary. Carrying 5 decimal places, calculate the regression coefficients. Ans. 0.0292, 0.6151, -0.2022.

EXAMPLE 13.10.8—Test the significance of the overall regression and compute the value of R^2 . Ans. $F = 16.2$, $f = 3$ and 40. P very small. $R^2 = 0.55$, a disappointing value when the objective is high accuracy in predicting Y .

EXAMPLE 13.10.9—Test the significance of the joint effect of X_1 and X_3 after fitting X_2 . Ans. $F = 0.87$. Can you conclude anything about the relative usefulness of the three predictors?

13.11—Numerical example. Computing the inverse matrix. Table 13.11.1 gives the worksheet in which the c 's are computed. The computing Instructions (column 2) are the same as in sections 13.9 and 13.10. The following points are worth noting:

1. In many problems the c 's are small numbers with numerous zeros after the decimal place. For those who have difficulty in keeping track of the zeros the following pre-coding is recommended. Code each X_i , if necessary, so that every Σx_i^2 lies between 0.1 and 10. This can always be done by dividing X_i by a power of 10. If X_1 is divided by 10^p and X_2 by 10^r , then Σx_1^2 is divided by 10^{2p} ; Σx_2^2 by 10^{2r} ; and $\Sigma x_1 x_2$ by 10^{p+r} . In this example we had initially (table 13.10.2), $\Sigma x_1^2 = 1,752.96$, $\Sigma x_2^2 = 3,155.78$, $\Sigma x_3^2 = 35,572.00$. Division of every X_i by 10^2 make the first two sums of squares lie between 0.1 and 1, while Σx_3^2 lies between 1 and 10 as shown in table 13.11.1. Every $\Sigma x_i x_j$ is also divided by 10^4 . The advantage is that the coded c 's are usually not far from 1. Five decimal places will be carried throughout the calculations.

2. The three sets of c 's are found simultaneously. The computations in column 6 give c_{11}, c_{12}, c_{13} , those in column 7 give c_{12}, c_{22}, c_{23} , and those in column 8 give c_{13}, c_{23}, c_{33} . Because of the symmetry, quantities like c_{12} are found only once.

3. Column 9, the sum of columns 3 to 8, is a check sum. Since mistakes creep in, check in each line indicated by a $\sqrt{\quad}$ that column 9 is the sum of columns 3 to 8. In some lines, e.g. (6), this check does not apply because of abbreviations in the method.

4. The first three numbers found in line (12) are c_{13}, c_{23}, c_{33} in coded form. Then we return to line (8). With column 7 as the right side, line (8) reads

$$c_{22} + 1.05533c_{23} = 4.02658$$

With column 6 as the right side, line (8) reads

$$c_{12} + 1.05533c_{13} = -2.49358$$

These give c_{22} and c_{12} . Finally, c_{11} comes from line (5).

5. To decode, c_{ij} is divided by the same factor by which $\Sigma x_i x_j$ was divided in coding.

6. By copying the $\Sigma x_i y$ next to the c_{ij} , the b_i are easily computed. Then the reduction in $S.S.$ due to regression and the Deviations mean square are obtained. These enable the standard error of each b_i to be placed next to b_i . As anticipated, neither b_2 nor b_3 approaches the significance level.

Occasionally there are several Y -variables whose sample regressions on the same set of X -variables are to be worked out. In the phosphorus

TABLE 13.11.1
CALCULATION OF INVERSE MATRIX. PLANT-AVAILABLE PHOSPHORUS

Line	1 Reciprocal	2 Instructions	3 X_1	4 X_2	5 X_3	6 c_{1j}	7 c_{2j}	8 c_{3j}	9 Check Sum
(1)			0.17530	0.10856	0.12000	1	0	0	1.40386
(2)			0.10856	0.31558	0.33640	0	1	0	1.76054
(3)			0.12000	0.33640	3.55720	0	0	1	5.01360
(4)		Copy (1)	0.17530	0.10856	0.12000	1	0	0	1.40386
(5)	5.70451	(4) \times 5.70451	1	[0.61928]	{0.68454}	5.70451	0	0	8.00833 ✓
(6)		(4) \times 0.61928		0.06723	0.07431	0.61928	0	0	0.86938
(7)		(2) - (6)		0.24835	0.26209	-0.61928	1	0	0.89116 ✓
(8)	4.02658	(7) \times 4.02658		1	{1.05533}	-2.49358	4.02658	0	3.58833 ✓
(9)		(4) \times [0.68454]			0.08214	0.68454	0	0	0.96099
(10)		(7) \times {1.05533}			0.27659	-0.65354	1.05533	0	0.94047
(11)		(3) - (9) - (10)			3.19847	-0.03100	-1.05533	-1	3.11214 ✓
(12)	0.31265	(11) \times 0.31265		1		c_{13}	c_{23}	c_{33}	0.97301 ✓
						-0.00969	-0.32995	0.31265	
		Line (8)	$c_{22} = 4.02658 - (1.05533)c_{23} = 4.37479$						
		Line (8)	$c_{12} = -2.49358 - (1.05533)c_{13} = -2.48335$						
		Line (5)	$c_{11} = 5.70451 - (0.61928)c_{12} - (0.68454)c_{13} = 7.24903$						
		Decoded c 's		$\Sigma x_i y_i$		b_i		$s_b = \sqrt{(s^2 c_{ii})}$	
c_{11}, c_{12}, c_{13}	0.000724903	-0.000248335	-0.000000969	3,231.48	1.7847**	0.538			
c_{21}, c_{22}, c_{23}	-0.000248335	0.000437479	-0.000032995	2,216.44	-0.0834	0.418			
c_{31}, c_{32}, c_{33}	-0.000000969	-0.000032995	0.000031265	7,593.00	0.1611	0.112			

$\Sigma y^2 = 12,390$: Reduction = $\Sigma b_i (\Sigma x_i y_i) = 6,806$ $s^2 = 5,584/14 = 399$

experiment, corn was grown in every soil sample at 35°C. as well as at 20°C. The amounts of phosphorus in the plants, Y' , are shown in the last column of table 13.10.1. Since the inverse matrix is the same for Y' as for Y , it is necessary to calculate only the new sums of products,

$$\Sigma x_1 y' = 1,720.42, \Sigma x_2 y' = 4,337.56, \Sigma x_3 y' = 8,324.00$$

Combining these with the c 's already calculated, the regression coefficients for Y' are

$$b_1' = 0.1619, b_2' = 1.1957, b_3' = 0.1155$$

In the new data, $\Sigma \hat{y}'^2 = 6,426$, $\Sigma d^2 = 12,390 - 6,426 = 5,964$, $s'^2 = 426.0$. The standard errors of the three regression coefficients are 0.556, 0.431, and 0.115. These lead to the three values of t : 0.29, 2.77, and 1.00. At 35°C., b_2 is the only significant regression coefficient. The interpretation made was that at 35°C. there was some mineralization of the organic phosphorus which would make it available to the plants.

The formulas for the standard errors of the estimates in multiple regression studies are illustrated in examples 13.11.1 to 13.11.3.

EXAMPLE 13.11.1—For soil sample 17, the predicted \hat{Y} was 119 ppm, and the x_i were: $x_1 = 14.9$, $x_2 = 15.9$, $x_3 = 79$. Find 95% limits for the population mean μ of Y . Ans. The variance of \hat{Y} as an estimate of μ is

$$s^2 \left\{ \frac{1}{n} + \Sigma c_{ii} x_i^2 + 2 \Sigma c_{ij} x_i x_j \right\}$$

The expression in the c 's is conveniently computed as follows:

	c_{ij}			x_i	$\Sigma c_{ij} x_j$
	.0007249	—	.0002483	—	.0000010
	—	.0002483	.0004375	—	.0000330
	—	.0000010	—	.0000330	.0000313
x_j	14.9	15.9	79.0		$\Sigma \Sigma c_{ij} x_i x_j = 0.2640$

Border the c_{ij} matrix with a row and a column of the x 's. Multiply each row of the c_{ij} in turn by the x_j , giving the sums of products 0.006774, etc. Then multiply this column by the x_i , giving the sum of products 0.2640. Since $n = 18$ and $s^2 = 399$, this gives

$$s_{\hat{Y}}^2 = (399)(0.0556 + 0.2640) = 127.5 \quad s_{\hat{Y}} = 11.3$$

With $t_{0.05} = 2.145$, the limits are $119 \pm (2.145)(11.3)$; 95 to 143 ppm.

EXAMPLE 13.11.2—If we are estimating Y for an individual new observation, the standard error of the estimate \hat{Y} is

$$\sqrt{s^2 \left\{ 1 + \frac{1}{n} + \Sigma c_{ii} x_i^2 + 2 \Sigma c_{ij} x_i x_j \right\}}$$

Verify that for a soil with the X -values of soil 17, the s.e. would be ± 22.9 ppm.

EXAMPLE 13.11.3—The following data, kindly provided by Dr. Gene M. Smith, come from a class of 66 students of nursing. Y represents the students' score in an examination on theory, X_1 the rank in high school (a high value being good), X_2 the score on a verbal aptitude test, and X_3 a measure of strength of character. The sums of squares and products (65 d.f.) are as follows:

$\Sigma x_i x_j$			$\Sigma x_i y$	Σy^2
24,633	2,212	5,865	925.3	670.3
	7,760	2,695	745.9	
		28,432	1,537.8	

(i) Show that the regression coefficients and their standard errors are as follows:

$$b_1 = 0.0206 \pm 0.0192; \quad b_2 = 0.0752 \pm 0.0340, \quad b_3 = 0.0427 \pm 0.0180$$

Which X variables are related to performance in theory?

(ii) Show that the F value for the three-variable regression is $F = 5.50$. What is the P value?

(ii) Verify that $R^2 = 0.210$.

13.12—Deletion of an independent variable. After a regression is computed, the utility of a variable may be questioned and its omission proposed. Instead of carrying out the calculations anew, the regression coefficients and the inverse matrix in the reduced regression can be obtained more quickly by the following formulas (14). We suppose that X_u is the variable to be omitted from a regression containing $X_1 \dots X_k$. Before omission, the Deviations mean square s^2 has $(n - k - 1)$ d.f.

When X_u is omitted, the sum of squares of deviations from the fitted regression, Σd^2 , is increased by b_u^2/c_{uu} . The mean square of the deviations then becomes

$$s'^2 = (\Sigma d^2 + b_u^2/c_{uu})/(n - k)$$

Further, the regression coefficients and the inverse multipliers become

$$b_i' = b_i - c_{iu}b_u/c_{uu}$$

$$c_{ii}' = c_{ii} - c_{iu}^2/c_{uu}$$

$$c_{ij}' = c_{ij} - c_{iu}c_{ju}/c_{uu}$$

13.13—Selection of variates for prediction. A related but more difficult problem arises when a regression is being constructed for purposes of prediction and it is thought that several of the X -variables, perhaps most of them, may contribute little or nothing to the accuracy of the prediction. For instance, we may start with 11 X -variables, but a suitable choice of three of them might give the best predictions. The problem is to decide how many variables to retain, and which ones.

The most thorough approach is to work out the regression of Y on every subset of the k X -variables, that is, on each variable singly, on

every pair of variables, on every triplet, and so on. The subset that gives the smallest Deviations mean square s^2 could be chosen, though if this subset involved 9 variables and another subset with 3 variables looked almost as good, the latter might be preferred for simplicity. The drawback of this method is the amount of computation. The number of regressions to be computed is $2^k - 1$, or 2,047 for 11 X -variables. Even with an electronic computer, this approach is scarcely feasible if k is large.

Two alternative approaches are the *step up method* and the *step down method*. In the *step down method*, the regression of Y on all k X -variables is calculated. The contribution of X_i to the reduction in sum of squares of Y , after fitting the other variables, is b_i^2/c_{ii} . The variable X_u for which this quantity is smallest is selected, and some rule is followed in deciding whether to omit X_u . One such rule is to omit X_u if $b_u^2/s^2c_{uu} < 1$: others omit X_u if b_u is not significant at some chosen level. If X_u is omitted, the regression of Y on the remaining $(k - 1)$ variables is computed, and the same rule is applied. The process continues until no variable qualifies for omission.

In the *step up method* we start with the regressions of Y on X_1, \dots, X_k taken singly. The variable giving the greatest reduction in sum of squares of Y is selected. Call this X_1 . Then the bivariate regressions in which X_1 appears are worked out. The variate which gives the greatest additional reduction in sum of squares after fitting X_1 is selected. Call this X_2 . All trivariate regressions that include both X_1 and X_2 are computed, and the variate that makes the greatest additional contribution to them is selected, and so on until this additional contribution b_i^2/c_{ii} is too small to satisfy some rule for inclusion.

It is known that the step up and the step down methods will not necessarily select the same X -variables, and that neither method guarantees to find the same variables as the exhaustive method of investigating every subset. Striking differences appear mainly when the X -variables are highly correlated. The differences are not necessarily alarming, because when intercorrelations are high, different subsets can give almost equally good predictions. Fuller accounts of these methods, with illustrations, appear in (15, 16).

Two aspects of this problem require further research. For a given approach, e.g., the step down method, the best rule to use in deciding whether to omit an X -variate is not clear. Naturally, all simple rules reject X_i if at some stage b_i^2/c_{ii} is small enough. Suppose that $\beta_i = +1$. Then X_i may be rejected because this sample gave an unusually low estimate of b_i , say 0.3. Nevertheless, with $\beta_i = +1$ a prediction formula that includes a term $0.3X_i$ may give better predictions in the population than one which has no term in X_i . For this reason some writers recommend retaining the term in X_i if the investigator is confident from his knowledge of the mechanism involved that β_i must be positive and if b_i is also positive.

Secondly, these methods tend to select variables that happen to do unusually well in the sample. When applied to new material, a prediction

formula selected in this way will not predict as accurately as the value of s^2 suggests, especially if the sample is small and many X 's have been rejected. More information is needed on the extent of this loss of accuracy.

13.14—The discriminant function. This is a multivariate technique for studying the extent to which different populations overlap one another or diverge from one another. It has three principal types of use.

1. *Classification and diagnosis.* The doctor's records of a person's symptoms and of his physical and laboratory measurements are taken to guide the doctor as to the particular disease from which the person is suffering. With two diseases that are often confused, it is helpful to learn what measurements are most effective in distinguishing between the conditions, how best to combine the measurements, and how successfully the distinction can be made.

2. *In the study of the relations between populations.* For example, to what extent do the aptitudes and attitudes of a competent architect differ from those of a competent engineer or a competent banker? Do non-smokers, cigarette smokers, pipe smokers, and cigar smokers differ markedly or only negligibly in their psychological traits?

3. *As a multivariate generalization of the t -test.* Given a number of related measurements made on each of two groups, the investigator may want a *single* test of the null hypothesis that the two populations have the same means with respect to all the measurements.

Historically, it is interesting that the discriminant function was developed independently by Fisher (17), whose primary interest was in classification, by Mahalanobis (18), in connection with a large study of the relations between Indian castes and tribes, and by Hotelling (19), who produced the multivariate t -test.

This introduction is confined to the case of two populations. Consider first a single variate X , normally distributed, with known means μ_1 , μ_2 in the two populations and known standard deviation σ , assumed the same in both populations. The value of X is measured for a new specimen that belongs to one of the two populations. Our task is to classify the specimen into the correct population. If $\mu_1 < \mu_2$, a natural classification rule is to assign the specimen to population *I* if $X < (\mu_1 + \mu_2)/2$, and to population *II* if $X > (\mu_1 + \mu_2)/2$. The mean of the two populations serves as the boundary point.

How often will we make a mistake? If the specimen actually comes from population *I*, our verdict is wrong whenever $X > (\mu_1 + \mu_2)/2$; that is, whenever

$$\frac{X - \mu_1}{\sigma} > \frac{\frac{\mu_1 + \mu_2}{2} - \mu_1}{\sigma} = \frac{(\mu_2 - \mu_1)}{2\sigma} = \frac{\delta}{2\sigma}$$

where $\delta = (\mu_2 - \mu_1)$ is the distance between the two means.

Equations (13.15.1) obviously resemble the normal equations for the regression coefficients in multiple regression. The L_i take the place of the b_i , and the d_i of the Σx_{ij} . The resemblance can be increased by constructing a dummy variable Y , which has the value $+1/n_2$ for every member of sample 2 and $-1/n_1$ for every member of sample 1. It follows that

$\Sigma X_i Y = \Sigma x_{iy} = d_i$. Thus, formally, the discriminant function can be regarded as the multiple regression of this dummy Y on $X_1 \dots X_k$. If we knew Y for any specimen we would know the population to which the specimen belongs. Consequently, it is reasonable that the discriminant function should try to predict Y as accurately as possible.

For the two sets of soils the normal equations are:

$$1.111L_1 + 0.229L_2 + 0.198L_3 = 0.1408$$

$$0.229L_1 + 1.043L_2 + 0.051L_3 = 0.0821$$

$$0.198L_1 + 0.051L_2 + 2.942L_3 = 0.0826$$

The L_i , computed by the method of section 13.10, are:

$$L_1 = 0.11229, \quad L_2 = 0.05310, \quad L_3 = 0.01960$$

The value of d/s for the discriminant is given by the formula:

$$\sqrt{(n_1 + n_2 - 2)\Sigma L_i d_i} = \sqrt{(284)(0.02179)} = \sqrt{6.188} = 2.49$$

This gives an estimated probability of misclassification of 10.6%. In these data the combined discriminant is not much better than pH alone.

TABLE 13.15.1
ANALYSIS OF VARIANCE OF THE DISCRIMINANT FUNCTION. HOTELLING'S T^2 -TEST

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Between soils	3	$n_1 n_2 (\Sigma L d)^2 / (n_1 + n_2) = 0.03088$	0.01029
Within soils	282	$\Sigma L d = 0.02179$	0.0000773

$$F = 0.01029/0.0000773 = 133.1.$$

The multivariate t -test, Hotelling's T^2 test, is made in table 13.15.1 from an analysis of variance of the variate $\Sigma L_i X_i$ into "Between Samples" and "Within Samples." On multiplying equations (13.15.1) by L_1, L_2, \dots, L_k and adding, we have the result:

$$\text{Within Samples sum of squares} = \Sigma \Sigma L_i L_j S_{ij} = \Sigma L_i d_i$$

$$\text{The "Between Samples" sum of squares} = n_1 n_2 (\Sigma L_i d_i)^2 / (n_1 + n_2)$$

Note the $d.f.$: k for Between Samples and $(n_1 + n_2 - k - 1)$ for Within Samples. The allocation of k $d.f.$ to Between Samples allows for the fact that the L 's were chosen to maximize the ratio of the Between Samples $S.S.$ to the Within Samples $S.S.$ The value of F , 133.1, with 3 and 282 $d.f.$ is very large, as it must be if the discriminant is to be effective in classification.

The assumption that the covariance matrix is the same in both populations is rather sweeping. If there appear to be moderate differences

between the matrices in the two populations and if n_1 and n_2 are unequal, it is better when computing the coefficients L_i to replace the sums of squares and products S_{ij} by the unweighted averages s_{ij} of the variances or covariances in the two samples. If this is done, note that the value of d/s for the discriminant becomes $\sqrt{\Sigma(Ld)}$, while in table 13.15.1, $\Sigma(Ld)$ becomes the Within Samples *mean square*. The expression for the Between Samples sum of squares remains as in table 13.15.1. When the covariance matrices differ substantially, the best discriminant is a quadratic expression in the X 's. Smith (23) presents an example of this case.

For classification studies involving more than two populations, see Rao (20). Examples are given in (24, 25) for qualitative data, in which the assumption of a multivariate normal population does not apply.

REFERENCES

1. M. T. EID, C. A. BLACK, O. KEMPTHORNE, and J. A. ZOELLNER. *Iowa Agric. Exp. Sta. Res. Bul.* 406 (1954).
2. R. A. FISHER. *Philos. Trans.*, B 213:89 (1924).
3. N. V. SMIRNOV. *Tables for the Distribution and Density Functions of t -distribution*. Pergamon, New York (1961).
4. G. E. P. BOX. *Technometrics*, 8:625 (1966).
5. O. KEMPTHORNE. *Design and Analysis of Experiments*. Wiley, New York (1952).
6. R. A. FISHER. *The Design of Experiments*. Oliver and Boyd, Edinburgh (1936).
7. H. HOTELLING. *Ann. Math. Statist.*, 11:271 (1940).
8. E. J. WILLIAMS. *Regression Analysis*. Wiley, Inc. New York (1959).
9. A. M. MOOD and F. A. GRAYBILL. *Introduction to the Theory of Statistics*. 2nd ed., McGraw-Hill, New York (1963).
10. M. H. DOOLITTLE. *U.S. Coast and Geodetic Survey Report* 115 (1878).
11. P. S. DWYER. *Linear Computations*. Wiley, New York (1951).
12. P. P. SWANSON, R. LEVERTON, M. R. GRAM, H. ROBERTS, and I. PESEK. *J. Gerontology*, 10:41 (1955).
13. A. M. BRUNSON and J. G. WILLIER. *J. Amer. Soc. Agron.*, 21:912 (1929).
14. W. G. COCHRAN. *J. R. Statist. Soc. Supp.*, 5:171 (1938).
15. N. DRAPER and H. SMITH. *Applied Regression Analysis*. Wiley, New York, Chap. 6 (1966).
16. H. C. HAMAKER. *Statist. Neerlandica*, 16:31 (1962).
17. R. A. FISHER. *Ann. Eugenics*, 7:179 (1936).
18. P. C. MAHALANOBIS. *J. Asiatic Soc. Bengal*, 26:541 (1930).
19. H. HOTELLING. *Ann. Math. Statist.*, 2:360 (1931).
20. C. R. RAO. *Advanced Statistical Methods in Biometric Research*. Wiley, New York, Chap. 8 (1952).
21. W. G. COCHRAN. *Technometrics*, 6:179 (1964).
22. G. M. COX and W. P. MARTIN. *Iowa State College Jour. Sci.*, 11:323 (1937).
23. C. A. B. SMITH. *Biomathematics*. Charles Griffin, London. (1954).
24. W. G. COCHRAN and C. E. HOPKINS. *Biometrics*, 17:10 (1961).
25. A. E. MAXWELL. *Analysing Qualitative Data*. Methuen, London. Chap. 10. (1961).
26. S. WRIGHT. *Biometrics*, 16:189 (1960).
27. O. D. DUNCAN. *Amer. J. Sociol.*, 72:1 (1966).
28. R. A. GROUT, *Iowa Agric. Exp. Sta. Res. Bul.*, 218 (1937).

Analysis of covariance

14.1—Introduction. The analysis of covariance is a technique that combines the features of analysis of variance and regression. In a one-way classification, the typical analysis of variance model for the value Y_{ij} of the j th observation in the i th class is

$$Y_{ij} = \mu_i + e_{ij}$$

where the μ_i represent the population means of the classes and the e_{ij} are the residuals. But suppose that on each unit we have also measured another variable X_{ij} that is linearly related to Y_{ij} . It is natural to set up the model,

$$Y_{ij} = \mu_i + \beta(X_{ij} - \bar{X}_{i..}) + \varepsilon_{ij},$$

where β is the regression coefficient of Y on X . This is a typical model for the analysis of covariance. If X and Y are closely related, we may expect this model to fit the Y_{ij} values better than the original analysis of variance model. That is, the residuals ε_{ij} should be in general smaller than the e_{ij} .

The model extends easily to more complex situations. With a two-way classification, as in a randomized blocks experiment, the model is

$$Y_{ij} = \mu + \alpha_i + \rho_j + \beta(X_{ij} - \bar{X}_{i..}) + \varepsilon_{ij}$$

With a one-way classification and two auxiliary variables X_{1ij} and X_{2ij} , both linearly related to Y_{ij} , we have

$$Y_{ij} = \mu_i + \beta_1(X_{1ij} - \bar{X}_{1i..}) + \beta_2(X_{2ij} - \bar{X}_{2i..}) + \varepsilon_{ij}$$

The analysis of covariance has numerous uses.

1. *To increase precision in randomized experiments.* In such applications the covariate X is a measurement, taken on each experimental unit before the treatments are applied, that predicts to some degree the final response Y on the unit. In the earliest application suggested by Fisher (1), the Y_{ij} were the yields of tea bushes in an experiment. An important

source of error is that by the luck of the draw, some treatments will have been allotted to a more productive set of bushes than others. The X_{ij} were the previous yields of the bushes in a period before treatments were applied. Since the relative yields of tea bushes show a good deal of stability from year to year, the X_{ij} serve as predictors of the inherent yielding abilities of the bushes. By adjusting the treatment mean yields so as to remove these differences in yielding ability, we obtain a lower experimental error and more precise comparisons among the treatments. This is probably the commonest use of covariance.

2. *To adjust for sources of bias in observational studies.* An investigator is studying the relation between obesity in workers and the physical activity required in their occupations. He has measures of obesity Y_{ij} in samples of workers from each of a number of occupations. He has also recorded the age X_{ij} of each worker, and notices that there are differences between the mean ages of the workers in different occupations. If obesity is linearly related to age, differences found in obesity among different occupations may be due in part to these age differences. Consequently he introduces the term $\beta(X_{ij} - \bar{X}..)$ into his model in order to adjust for a possible source of bias in his comparison among occupations.

3. *To throw light on the nature of treatment effects in randomized experiments.* In an experiment on the effects of soil fumigants on nematodes, which attack some farm crops, significant differences between fumigants were found both in the numbers of nematode cysts X_{ij} and in the yields Y_{ij} of the crop. This raises the question: Can the differences in yields be ascribed to the differences in numbers of nematodes? One way of examining this question is to see whether treatment differences in yields remain, or whether they shrink to insignificance, after adjusting for the regression of yields on nematode numbers.

4. *To study regressions in multiple classifications.* For example, an investigator is studying the relation between expenditure per student in schools (Y) and per capita income (X) in large cities. If he has data for a large number of cities for each of four years, he may want to examine whether the relation is the same in different sections of the country, or whether it remains the same from year to year. Sometimes the question is whether the relation is straight or curved.

14.2—Covariance in a completely randomized experiment. We begin with a simple example of the use of covariance in increasing precision in randomized experiments. With a completely randomized design, the data form a one-way classification, the treatments being the classes. In the model

$$Y_{ij} = \mu_i + \beta(X_{ij} - \bar{X}..) + \varepsilon_{ij},$$

the μ_i represent the effects of the treatments. The observed mean for the i th treatment is

$$\bar{Y}_i = \mu_i + \beta(\bar{X}_i - \bar{X}..) + \bar{e}_i.$$

Thus \bar{Y}_i is an unbiased estimate of

$$\mu_i + \beta(\bar{X}_i - \bar{X}..)$$

It follows that as an estimate of μ_i we use

$$\hat{\mu}_i = \bar{Y}_i - \beta(\bar{X}_i - \bar{X}..),$$

the second term on the right being the adjustment introduced by the covariance analysis. The adjustment accords with common sense. For instance, suppose we were told that in the previous year the tea bushes receiving Treatment 1 yielded 20 pounds more than the average over the experiment. If the regression coefficient of Y on X was 0.4, meaning that each pound of increase in X corresponds to 0.4 pound of increase in Y , we would decrease the observed Y mean by $(0.4)(20) = 8$ pounds in order to make Treatment 1 more comparable to the other treatments. In this illustration the figure 0.4 is β and the figure 20 is $(\bar{X}_i - \bar{X}..)$.

There remains the problem of estimating β from the results of the experiment. In a single sample you may recall that the regression coefficient is estimated by $b = \Sigma xy / \Sigma x^2$, and that the reduction in sum of squares of Y due to the regression is $(\Sigma xy)^2 / \Sigma x^2$. These results continue to hold in multiple classifications (completely randomized, randomized blocks and Latin square designs) except that β is estimated from the Error line in the analysis of variance. We may write $b = E_{xy} / E_{xx}$. The Error sum of squares of X in the analysis of variance, E_{xx} , is familiar, but the quantity E_{xy} is new. It is the Error sum of products of X and Y . A numerical example will clarify it.

The data in table 14.2.1 were selected from a larger experiment on the use of drugs in the treatment of leprosy at the Eversley Childs Sanitarium in the Philippines. On each patient six sites on the body at which leprosy bacilli tend to congregate were selected. The variate X , based on laboratory tests, is a score representing the abundance of leprosy bacilli at these sites before the experiment began. The variate Y is a similar score after several months of treatment. Drugs A and D are antibiotics while drug F is an inert drug included as a control. Ten patients were selected for each treatment for this example.

The first step is to compute the analysis of sums of squares and products, shown under the table. In the columns headed Σx^2 and Σy^2 , we analyze X and Y in the usual way into "Between drugs" and "Within drugs." For the Σxy column, make the corresponding analysis of the products of X and Y , as follows:

$$\text{Total: } (11)(6) + (8)(0) + \dots + (12)(20) - (322)(237)/30 = 731.2$$

$$\text{Between drugs: } \frac{(93)(53) + (100)(61) + (129)(123)}{10} - \frac{(322)(237)}{30} = 145.8$$

TABLE 14.2.1
SCORES FOR LEPROSY BACILLI BEFORE (X) AND AFTER (Y) TREATMENT

Drugs							
A		D		F			
X	Y	X	Y	X	Y		
11	6	6	0	16	13		
8	0	6	2	13	10		
5	2	7	3	11	18		
14	8	8	1	9	5		
19	11	18	18	21	23		
6	4	8	4	16	12		
10	13	19	14	12	5		
6	1	8	9	12	16		
11	8	5	1	7	1		
3	0	15	9	12	20		
Totals	93	53	100	61	129	123	
Means	9.3	5.3	10.0	6.1	12.9	12.3	
							Overall
							\bar{X} \bar{Y}
							322 237
							10.73 7.90

Analysis of Sums of Squares and Products

Source	$d.f.$	Σx^2	Σxy	Σy^2
Total	29	665.9	731.2	1,288.7
Between drugs	2	73.0	145.8	293.6
Within drugs (Error)	27	592.9	585.4	995.1
Reduction due to regression	1		$(585.4)^2/592.9 = 578.0$	
Deviations from regression	26			417.1

$$\text{Deviations mean square} = 417.1/26 = 16.04$$

The Within drugs sum of products, 585.4, is found by subtraction. Note that any of these sums of products may be either positive or negative. The Within drugs (Error) sum of products 585.4 is the quantity we call E_{xy} , while the Error sum of squares of X , 592.9, is E_{xx} .

The reduction in the Error sum of squares Y due to the regression is E_{xy}^2/E_{xx} with 1 $d.f.$ The Deviations mean square, 16.04 with 26 $d.f.$, provides the estimate of error. The original Error mean square of Y is $995.1/27 = 36.86$. The regression has produced a substantial reduction in the Error mean square.

The next step is to compute b and the adjusted means. We have $b = E_{xy}/E_{xx} = 585.4/592.9 = 0.988$. The adjusted means are as follows:

$$A: \bar{Y}_1 - b(\bar{X}_1 - \bar{X}_{..}) = 5.3 - (0.988)(9.3 - 10.73) = 6.71$$

$$D: \bar{Y}_2 - b(\bar{X}_2 - \bar{X}_{..}) = 6.1 - (0.988)(10.0 - 10.73) = 6.82$$

$$F: \bar{Y}_3 - b(\bar{X}_3 - \bar{X}_{..}) = 12.3 - (0.988)(12.9 - 10.73) = 10.16$$

have improved the status of F , which happened to receive initially a set of patients with somewhat high scores.

For tests of significance or confidence limits relating to the adjusted means, the error variance is derived from the mean square $s_{y \cdot x}^2 = 16.04$, with 26 *df*. Algebraically, the difference between the adjusted means of the *i*th and the *j*th treatments is

$$D = \bar{Y}_i - \bar{Y}_j - b(\bar{X}_i - \bar{X}_j)$$

The formula for the estimated variance of *D* is

$$s_D^2 = s_{y \cdot x}^2 \left\{ \frac{2}{n} + \frac{(\bar{X}_i - \bar{X}_j)^2}{E_{xx}} \right\} \quad (14.2.1)$$

where *n* is the sample size per treatment. The second term on the right is an allowance for the sampling error of *b*.

This formula has the disadvantage that s_D is different for every pair of treatments that are being compared. In practice, these differences are small if (i) there are at least 20 *df* in the Error line of the analysis of variance, and (ii) the Treatments mean square for *X* is non-significant, as it should be since the *X*'s were measured before treatments were assigned. In such cases an average value of s_D^2 may be used. By an algebraic identity (2) the average value of s_D^2 , taken over every pair of treatments, is

$$\overline{s_D^2} = \frac{2}{n} s_{y \cdot x}^2 \left[1 + \frac{t_{xx}}{E_{xx}} \right] \quad (14.2.2)$$

where t_{xx} is the Treatments mean square for *X*. More generally, we may regard

$$s'^2 = s_{y \cdot x}^2 \left[1 + \frac{t_{xx}}{E_{xx}} \right] \quad (14.2.3)$$

as the effective Error mean square per observation when computing the error variance for any comparison among the treatment means.

In this experiment $t_{xx} = 73.0/2 = 36.5$ (from table 14.2.1), $E_{xx} = 592.9$, giving $t_{xx}/E_{xx} = 0.0616$. Hence,

$$s'^2 = (16.04)(1.0616) = 17.03 \quad : \quad s' = 4.127$$

With 10 replicates this gives $\overline{s_D} = 4.127/\sqrt{(0.2)} = 1.846$. The adjusted means for *A* and *D*, 6.71 and 6.82, show no sign of a real difference. The largest contrast, *F* - *A*, is 3.45, giving a *t*-value of $3.45/1.846 = 1.87$, with 26 *df*., which is not significant at the 5% level.

After completing a covariance analysis, the experimenter is sure to ask: Is it worthwhile? The efficiency of the adjusted means relative to the unadjusted means is estimated by the ratio of the corresponding effective Error mean squares:

$$\frac{s_y^2}{s'^2} = \frac{s_y^2}{s_{y \cdot x}^2 \left[1 + \frac{t_{xx}}{E_{xx}} \right]} = \frac{36.86}{17.03} = 2.16$$

Covariance with 10 replicates per treatment gives nearly as precise estimates as the unadjusted means with 21 replicates.

In experiments like this, in which X measures the same quantity as Y (score for leprosy bacilli), an alternative to covariance is to use $(Y - X)$, the *change* in the score, as the measure of treatment effect. The Error mean square for $(Y - X)$ is obtained from table 14.2.1 as

$$\frac{E_{yy} - 2E_{xy} + E_{xx}}{27} = \frac{[995.1 - 2(585.4) + 592.9]}{27} = 15.45$$

This compares with 17.03 for covariance. In this experiment, use of $(Y - X)$ is slightly more efficient than covariance as well as quicker computationally. This was the recommended variable for analysis in the larger experiment from which these data were selected. In many experiments, $(Y - X)$ is inferior to covariance, and may also be inferior to Y if the correlation between X and Y is low.

14.3—The F -test of the adjusted means. Section 14.2 has shown how to make comparisons among the adjusted means. It is also possible to perform an F -test of the null hypothesis that all the μ_i are equal—that there are no differences among the adjusted means. Since the way in which this test is computed often looks mystifying, we first explain its rationale.

First we indicate why b is always estimated from the Error line of the analysis of variance. Suppose that the value of b has not yet been chosen. As we have seen, the analysis of covariance is essentially an analysis of variance of the quantity $(Y - bX)$. The Error sum of squares of this quantity may be written

$$E_{yy} - 2bE_{xy} + b^2E_{xx}$$

Completing the square on b , the Error S.S. is

$$E_{xx} \left(b - \frac{E_{xy}}{E_{xx}} \right)^2 + E_{yy} - \frac{E_{xy}^2}{E_{xx}} \quad (14.3.1)$$

By the method of least squares, the value of b is selected so as to minimize the Error S.S. From (14.3.1), it is obvious that this happens when $b = E_{xy}/E_{xx}$, the minimum Error S.S. being $E_{yy} - E_{xy}^2/E_{xx}$.

Now to the F -test. If the null hypothesis is true, a covariance model in which $\mu_i' = \mu$ should fit the data as well as the original covariance model. Consequently, we fit this H_0 model to find how large an Error S.S. it gives. In the analysis of sums of squares and products for the H_0 model, the "Error" line is the sum of the Error and Treatments line in the original model, because the H_0 model contains no treatment effects. Hence, the Deviations S.S. from the H_0 model is

$$E_{yy} + T_{yy} - \frac{(E_{xy} + T_{xy})^2}{E_{xx} + T_{xx}} \quad (14.3.2)$$

If H_0 holds, the difference between the Deviations S.S. for the H_0 model and the original model, when divided by the difference in degrees of freedom, may be shown to be an estimate of $\sigma_{y \cdot x}^2$ in the original model. If H_0 is false, this mean square difference becomes large because the H_0 model fits poorly. This mean square difference forms the numerator of the F -test. The denominator is the Deviations mean square from the original model.

In table 14.3.1 the test is made for the leprosy example. The first step is to form a Treatments + Error line. (In a completely randomized design this line is, of course, the same as the Total line, but this is not so in randomized blocks or a Latin square.) Following formula (14.3.2) we subtract $(731.2)^2/665.9 = 802.9$ from 1288.7 to give the deviations S.S., 485.8, for the H_0 model. From this we subtract 417.1, the deviations S.S. for the original model, and divide by the difference in $d.f.$, 2. The F -ratio, $34.35/16.04 = 2.14$, with 2 and 26 $d.f.$, lies between the 25% and the 10% levels.

TABLE 14.3.1
THE COVARIANCE F -TEST IN A ONE-WAY CLASSIFICATION. LEPROSY DATA

						Deviations From Regression		
	Degrees of Freedom	Σx^2	Σxy	Σy^2	Red	Degrees of Freedom	Sum of Squares	Mean Square
Treatments	2	73 0	145 8	293 6	578.0	26	417.1	16.04
Error	27	592 9	585.4	995 1				
$T + E$	29	665 9	731.2	1,288.7	802.9	28	485.8	
						2	68.7	34.35

14.4—Covariance in a two-way classification. The computations involve nothing new. The regression coefficient is estimated from the Error (Treatments \times Blocks) line in the analysis of sums of squares and products, and the F -test of the adjusted treatment means is made by recomputing the regression from the Treatments plus Error lines, following the procedure in section 14.3. To put it more generally for applications in which the words "Treatments" and "Blocks" are inappropriate, the regression coefficient is estimated from the Rows \times Columns line, and either the adjusted row means or the adjusted column means may be tested. Two examples from experiments will be presented to illustrate points that arise in applications.

The data in table 14.4.1 are from an experiment on the effects of two drugs on mental activity (13). The mental activity score was the sum of the scores on seven items in a questionnaire given to each of 24 volunteer subjects. The treatments were morphine, heroin, and placebo (an inert substance), given in subcutaneous injections. On different occasions, each

TABLE 14.4.1
MENTAL ACTIVITY SCORES BEFORE (X) AND TWO HOURS AFTER (Y) A DRUG

Subject	Morphine		Heroin		Placebo		Total	
	X	Y	X	Y	X	Y	X	Y
1	7	4	0	2	0	7	7	13
2	2	2	4	0	2	1	8	3
3	14	14	14	13	14	10	42	37
4	14	0	10	0	5	10	29	10
5	1	2	4	0	5	6	10	8
6	2	0	5	0	4	2	11	2
7	5	6	6	1	8	7	19	14
8	6	0	6	2	6	5	18	7
9	5	1	4	0	6	6	15	7
10	6	6	10	0	8	6	24	12
11	7	5	7	2	6	3	20	10
12	1	3	4	1	3	8	8	12
13	0	0	1	0	1	0	2	0
14	8	10	9	1	10	11	27	22
15	8	0	4	13	10	10	22	23
16	0	0	0	0	0	0	0	0
17	11	1	11	0	10	8	32	9
18	6	2	6	4	6	6	18	12
19	7	9	0	0	8	7	15	16
20	5	0	6	1	5	1	16	2
21	4	2	11	5	10	8	25	15
22	7	7	7	7	6	5	20	19
23	0	2	0	0	0	1	0	3
24	12	12	12	0	11	5	35	17
Total	138	88	141	52	144	133	423	273

	Degrees of Freedom	Σx^2	Σxy	Σy^2
Between subjects	23	910	519	558
Between drugs	2	1	5	137
Error	46	199	-16	422
Total	71	1,110	508	1,117

subject received each drug in turn. The mental activity was measured before taking the drug (X) and at 1/2, 2, 3, and 4 hours after. The response data (Y) in table 14.4.1 are those at two hours after. As a common precaution in these experiments, eight subjects took morphine first, eight took heroin first, and eight took the placebo first, and similarly on the second and third occasions. In these data there was no apparent effect of the order in which drugs were given, and the order is ignored in the analysis of variance presented here.

In planning this experiment two sources of variation were recognized. First, there are consistent differences in level of mental activity

between subjects. This source was removed from the experimental error by the device of having each subject test all three drugs, so that comparisons between drugs are made within subjects. Secondly, a subject's level changes from time to time—he feels sluggish on some occasions and unusually alert on others. Insofar as these differences are measured by the pretest mental activity score on each occasion, the covariance analysis should remove this source of error.

As it turned out, the covariance was ineffective in this experiment. The error regression coefficient is actually slightly negative, $b = -16/199$, and showed no sign of statistical significance. Consequently, comparison of the drugs is best made from the 2-hour readings alone in this case. Incidentally, covariance would have been quite effective in removing differences in mental activity between subjects, since the Between subjects b , 519/910, is positive and strongly significant.

Unlike the previous leprosy example, the use of the change in score, 2 hours – pretest, would have been unwise as a measure of the effects of the drugs. From table 14.4.1 the Error sum of squares for $(Y - X)$ is

$$422 + 199 - 2(-16) = 653$$

This is substantially larger than the sum of squares, 422, for Y alone.

The second example, table 14.4.2, illustrates another issue (3). The experiment compared the yields Y of six varieties of corn. There was some variation from plot to plot in number of plants (stand). If this variation is caused by differences in fertility in different plots and if higher plant numbers result in higher yields per plot, increased precision will be obtained by adjusting for the covariance of yield on plant number. The plant numbers in this event serve as an index of the fertility levels of the plots. But if some varieties characteristically have higher plant numbers than others through a greater ability to germinate or to survive when the plants are young, the adjustment for stand distorts the yields because it is trying to compare the varieties at some average plant number level that the varieties do not attain in practice.

With this in mind, look first at the F -ratio for Varieties in X (stand). From table 14.4.2 the mean squares are: Varieties 9.17, Error 7.59, giving $F = 1.21$. The low value of F gives assurance that the variations in stand are mostly random and that adjustment for stand will not introduce bias.

In the analysis, note the use of the Variety plus Error line in computing the F -test of the adjusted means. The value of F is $645.38/97.22 = 6.64$, highly significant with 5 and 14 $d.f.$ The adjustment produced a striking decrease in the Error mean square, from 583.5 to 97.2, and an increase in F from 3.25 to 6.64.

The adjusted means will be found to be:

$$A, 191.8; \quad B, 191.0; \quad C, 193.1; \quad D, 219.3; \quad E, 189.6; \quad F, 213.6$$

The standard error of the difference between two adjusted means is 7.25, with 14 $d.f.$ By either the LSD method or the sequential Newman-Keuls

TABLE 14.4.2
STAND (X) AND YIELD (Y) (POUNDS FIELD WEIGHT OF EAR CORN) OF SIX VARIETIES OF
CORN. COVARIANCE IN RANDOMIZED BLOCKS

Varieties	Blocks								Total	
	1		2		3		4			
	<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>
<i>A</i>	28	202	22	165	27	191	19	134	96	692
<i>B</i>	23	145	26	201	28	203	24	180	101	729
<i>C</i>	27	188	24	185	27	185	28	220	106	778
<i>D</i>	24	201	28	231	30	238	30	261	112	931
<i>E</i>	30	202	26	178	26	198	29	226	111	804
<i>F</i>	30	228	25	221	27	207	24	204	106	860
Total	162	1,166	151	1,181	165	1,222	154	1,225	632	4,794

					Deviations From Regression		
Source of Variation	$d.f.$	Σx^2	Σxy	Σy^2	$d.f.$	Sum of Squares	Mean Square
Total	23	181.33	1,485.00	18,678.50			
Blocks	3	21.67	8.50	436.17			
Varieties	5	45.83	559.25	9,490.00			
Error	15	113.83	917.25	8,752.33	14	1,361.07	97.22
Variety plus error	20	159.66	1,476.50	18,242.33	19	4,587.99	
For testing adjusted means,					5	3,226.92	645.38**

method, the two highest yielding varieties, *D* and *F*, are not significantly different, but they are significantly superior to all the others, which do not differ significantly among themselves.

In some cases, plant numbers might be influenced partly by fertility variations and partly by basic differences between varieties. The possibility of a partial adjustment has been considered by H. F. Smith (4).

EXAMPLE 14.4.1—Verify the adjusted means in the corn experiment and carry through the tests of all the differences.

EXAMPLE 14.4.2—Estimate the efficiency of the covariance adjustments. Ans. 5.55.

EXAMPLE 14.4.3—As an alternative to covariance, could we analyze the yield per plant, Y/X , as a means of removing differences in plant numbers? Ans. This is satisfactory if the relation between Y and X is a straight line going through the origin. But b is often substantially less than the mean yield per plant, because when plant numbers are high, competition between plants reduces the yield per plant. If this happens, the use of Y/X overcorrects for stand. In the corn example $b = 8.1$ and the overall yield per plant is $4,794/632 = 7.6$, in good agreement: Yield per plant would give results similar to covariance. Of course, yield per plant should be analyzed if there is direct interest in this quantity.

EXAMPLE 14.4.4—The following data are the yields (Y) in bushels per acre and the per cents of stem canker infection (X) in a randomized blocks experiment comparing four lines of soybeans (5).

Blocks	Lines								Totals	
	<i>A</i>		<i>B</i>		<i>C</i>		<i>D</i>			
	<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>
1	19.3	21.3	10.1	28.3	4.3	26.7	14.0	25.1	47.7	101.4
2	29.2	19.7	34.7	20.7	48.2	14.7	30.2	20.1	142.3	75.2
3	1.0	28.7	14.0	26.0	6.3	29.0	7.2	24.9	28.5	108.6
4	6.4	27.3	5.6	34.1	6.7	29.0	8.9	29.8	27.6	120.2
Totals	55.9	97.0	64.4	109.1	65.5	99.4	60.3	99.9	246.1	405.4

By looking at some plots with unusually high and unusually low X , note that there seems a definite negative relation between Y and X . Before removing this source of error by covariance, check that the lines do not differ in the amounts of infection. The analysis of sums of squares and products is as follows:

	<i>df.</i>	Σx^2	Σxy	Σy^2
Blocks	3	2,239.3	-748.0	272.9
Treatments	3	14.3	10.2	21.2
Error	9	427.0	-145.7	66.0
<i>T + E</i>	12	441.3	-135.5	87.2

- (i) Perform the F -test of the adjusted means.
(ii) Find the adjusted means and test the differences among them.
(iii) Estimate the efficiency of the adjustments. Ans. (i) $F = 4.79^*$; $df. = 3, 8$; (ii) $A, 23.77$; $B, 27.52$; $C, 25.19$; $D, 24.87$. By the LSD test, B significantly exceeds A and D .
(iii) 3.56. Strictly, a slight correction to this figure should be made for the reduction in $df.$ from 9 to 8.

14.5—Interpretation of adjusted means in covariance. The most straightforward use of covariance has been illustrated by the preceding examples. In these, the covariate X is a measure of the responsiveness of the experimental unit, either directly (as with the leprosy bacilli) or indirectly (as with number of plants). The adjusted means are regarded as better estimates of the treatment effects than the unadjusted means because one of the sources of experimental error has been removed by the adjustments.

Interpretation of adjusted means is usually more difficult when both Y and X show differences between treatments, or between groups in an observational study. As mentioned in section 14.1, adjusted means are sometimes calculated in this situation either in order to throw light on the way in which the treatments produce their effects or to remove a source of bias in the comparison of Y between groups. The computations remain unchanged, except that the use of the effective Error mean square

$$s_{y \cdot x}^2 \left[1 + \frac{t_{xx}}{E_{xx}} \right]$$

is not recommended for finding an approximation to the variance of the difference between two adjusted means. Instead, use the correct formula:

$$s_D^2 = s_{y \cdot x}^2 \left\{ \frac{2}{n} + \frac{(\bar{X}_{i \cdot} - \bar{X}_{j \cdot})^2}{E_{xx}} \right\}$$

The reason is that when the X 's differ from treatment to treatment, the term $(\bar{X}_{i \cdot} - \bar{X}_{j \cdot})^2$ can be large and can vary materially from one pair of means to another, so that s_D^2 is no longer approximately constant.

As regards interpretation, the following points should be kept in mind. If the X 's vary widely between treatments or groups, the adjustment involves an element of extrapolation. To cite an extreme instance, suppose that one group of men have ages (X) in the forties, with mean about 45, while a second group are in their fifties with mean about 55. In the adjusted means, the two groups are being compared at mean age 50, although neither group may have any men at this specific age. In using the adjustment, we are assuming that the linear relation between Y and X holds somewhat beyond the limits of each sample. In this situation the value of s_D^2 becomes large, because the term $(\bar{X}_{i \cdot} - \bar{X}_{j \cdot})^2$ is large. The formula is warning us that the adjustments have a high element of uncertainty. It follows that the comparison of adjusted means has low precision. Finding that F - or t -tests of the adjusted means show no significance, we may reach the conclusion that "The differences in Y can be explained as a consequence of the differences in X ," when a sounder interpretation is that the adjusted differences are so imprecise that only very large effects could have been detected. A safeguard is to compute confidence limits for some of the adjusted differences: if the F -test alone is made, this point can easily be overlooked.

Secondly, if X is subject to substantial errors of measurement, the adjustment removes only part of any difference between the Y means that is due to differences in the X means. Under the simplest mathematical model, the fraction removed may be shown to be $\sigma_X^2/(\sigma_X^2 + \sigma_d^2)$, where σ_d^2 is the variance of the errors of measurement of X . This point could arise in an example mentioned in section 14.1, in which covariance was suggested for examining whether differences produced by soil fumigants on spring oats (Y) could be explained as a reflection of the effects of these treatments on the numbers of nematode cysts (X). The nematode cysts are counted by taking a number of small soil samples from each plot and sifting each sample carefully by some process. The estimate of X on each plot is therefore subject to a sampling error and perhaps also to an error caused by failure to detect some of the cysts. Because of these errors, some differences might remain among the adjusted Y means, leading to an erroneous inference that the differences in yield could not be fully ex-

plained by the effects of the treatments on the nematodes. Similarly, in observational studies the adjustment removes only a fraction $\sigma_x^2/(\sigma_x^2 + \sigma_d^2)$ of a bias due to a linear relation between Y and X . Incidentally, the errors of measurement d do not vitiate the use of covariance in increasing the precision of the Y comparisons in randomized experiments, provided that Y has a linear regression on the measurement $X' = X + d$. However, as might be expected, they make the adjustments less effective, because the correlation ρ' between Y and $X' = X + d$ is less than the correlation ρ between Y and X , so that the residual error variance $\sigma_Y^2(1 - \rho'^2)$ is larger.

Finally, the meaning of the adjusted values is often hard to grasp, especially if the reasons for the relation between Y and X are not well known. As an illustration, table 14.5.1 shows the average 1964 expenditures \bar{Y} per attending pupil for schools in the states in each of five regions of the U.S. (6). These are simple averages of the values for the individual states in the region. Also shown are corresponding averages of 1963 per capita incomes \bar{X} in each region. In an analysis of variance into Between Regions and Between States Within Regions, the differences between regions are significant both for the expenditure figures and the per capita incomes. Further, the regions fall in the same order for expenditures as for incomes.

TABLE 14.5.1
1964 SCHOOL EXPENDITURES PER ATTENDING PUPIL (\bar{Y}) AND 1963 PER CAPITA INCOMES (\bar{X}) IN FIVE REGIONS OF THE U.S.

	East	Mountain and Pacific	North Central	South Atlantic	South Central
Number of states	8	11	12	9	8
	(dollars)				
Expenditures	542	500	479	399	335
Per capita incomes	2,600	2,410	2,370	2,310	1,780

It seems natural to ask: Would the differences in expenditures disappear after allowing for the relation between expenditure and income? The within-region regression appears to be linear, and the values of b do not differ significantly from region to region. The average b is 0.140 (\$14 in expenditure for each additional \$100 of income). The adjusted means for expenditure, adjusted to the overall average income of \$2,306, are as follows:

	<i>E</i>	<i>M P</i>	<i>N C</i>	<i>S A.</i>	<i>S C</i>
(Dollars)	501	485	470	398	409

The differences between regions have now shrunk considerably, although still significant, and the regions remain in the same order except that the South Central region is no longer lowest. On reflection, however, these adjusted figures seem hypothetical rather than concrete. The figure of \$409 for the South Central region cannot be considered an estimate of the amount that this region would spend per pupil if its per capita income were to increase rapidly, perhaps through greater industrialization, from \$1,780 to \$2,306. In fact, if we were trying to estimate this amount, a study of the Between Years regression of expenditure on income for individual states would be more relevant. Similarly, a conclusion that "the differences in expenditures cannot be ascribed to differences in per capita income" is likely to be misunderstood by a non-technical reader. For a good discussion of other complications in interpretation, see (4).

14.6—Comparison of regression lines. Frequently, the relation between Y and X is studied in samples obtained by different investigators, or in different environments, or at different times. In summarizing these results, the question naturally arises: can the regression lines be regarded as the same? If not, in what respects do they differ? A numerical example provides an introduction to the handling of these questions. The example has only two samples, but the techniques extend naturally to more than two samples.

In a survey to examine relationships between the nutrition and the health of women in the Middle West (7), the concentration of cholesterol in the blood serum was determined on 56 randomly selected subjects in Iowa and 130 in Nebraska. In table 14.6.1 are subsamples from the survey data. Figure 14.6.1 shows graphs of the data from each state. The figure gives an impression of linearity of the regression of cholesterol concentration on age, which will be assumed in this discussion.

The purpose is to examine whether the linear regressions of cholesterol on age are the same in Iowa and Nebraska. They may differ in slope, in elevation, or in the residual variances $\sigma_{y \cdot x}^2$. The most convenient approach is to compare the residual variances first, then the slopes, and lastly the elevations. In terms of the model, we have

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + \varepsilon_{ij},$$

where $i = 1, 2$ denotes the two states. We first compare the residual variances σ_1^2 and σ_2^2 , next β_1 and β_2 , and finally the elevations of the lines, α_1 and α_2 .

The computations begin by recording separately the Within sum of squares and products for each state, as shown in table 14.6.2 on lines 1 and 2. The next step is to find the residual $S.S.$ from regression for each state, as on the right in lines 1 and 2. The Residual mean squares, 2,392 and 1,581, are compared by the two-tailed F -test (section 2.9) or, with more than two samples, by Bartlett's test (section 10.21). If heterogeneous variances were evident, this might be pertinent information in itself. In

TABLE 1461
AGE AND CONCENTRATION OF CHOLESTEROL (MG /100 ML) IN THE BLOOD SERUM OF
IOWA AND NEBRASKA WOMEN

Iowa, $n = 11$		Nebraska, $n = 19$			
Age X	Cholesterol Y	Age X	Cholesterol Y	Age X	Cholesterol Y
46	181	18	137	30	140
52	228	44	173	47	196
39	182	33	177	58	262
65	249	78	241	70	261
54	259	51	225	67	356
33	201	43	223	31	159
49	121	44	190	21	191
76	339	58	257	56	197
71	224	63	337		
41	112	19	189		
58	189	42	214		
Sum 584	2,285			873	4,125
$\bar{X}_I = 53.1$	$\bar{Y}_I = 207.7$			$\bar{X}_N = 45.9$	$\bar{Y}_N = 217.1$

Iowa		
$\Sigma X^2 = 32,834$	$\Sigma XY = 127,235$	$\Sigma Y^2 = 515,355$
$C: 31,005$	121,313	474,657
$\Sigma x^2 = 1,829$	$\Sigma xy = 5,922$	$\Sigma y^2 = 40,698$

Nebraska		
$\Sigma X^2 = 45,677$	$\Sigma XY = 203,559$	$\Sigma Y^2 = 957,785$
$C: 40,112$	189,533	895,559
$\Sigma x^2 = 5,565$	$\Sigma xy = 14,026$	$\Sigma y^2 = 62,226$

Total, $n = 30$			
$\Sigma X = 1,457, \bar{X}_T = 48.6$	$\Sigma X^2 = 78,511$	$\Sigma XY = 330,794$	$\Sigma Y^2 = 1,473,140$
$\Sigma Y = 6,410, \bar{Y}_T = 213.7$	$C: 70,762$	311,312	1,369,603
	$\Sigma x^2 = 7,749$	$\Sigma xy = 19,482$	$\Sigma y^2 = 103,537$

this example, $F = 1.51$, with 9 and 17 *d.f.*, giving a P value greater than 0.40 in a two-tailed test. The mean squares show no sign of a real difference.

Assuming homogeneity of residual variances, we now compare the two slopes or regression coefficients, 3.24 for Iowa and 2.52 for Nebraska. A look at the scatters of the points about the individual regression lines in figure 14.6.1 suggests that the differences in slope may be attributable to sampling variation. To make the test (table 14.6.2), add the *d.f.* and

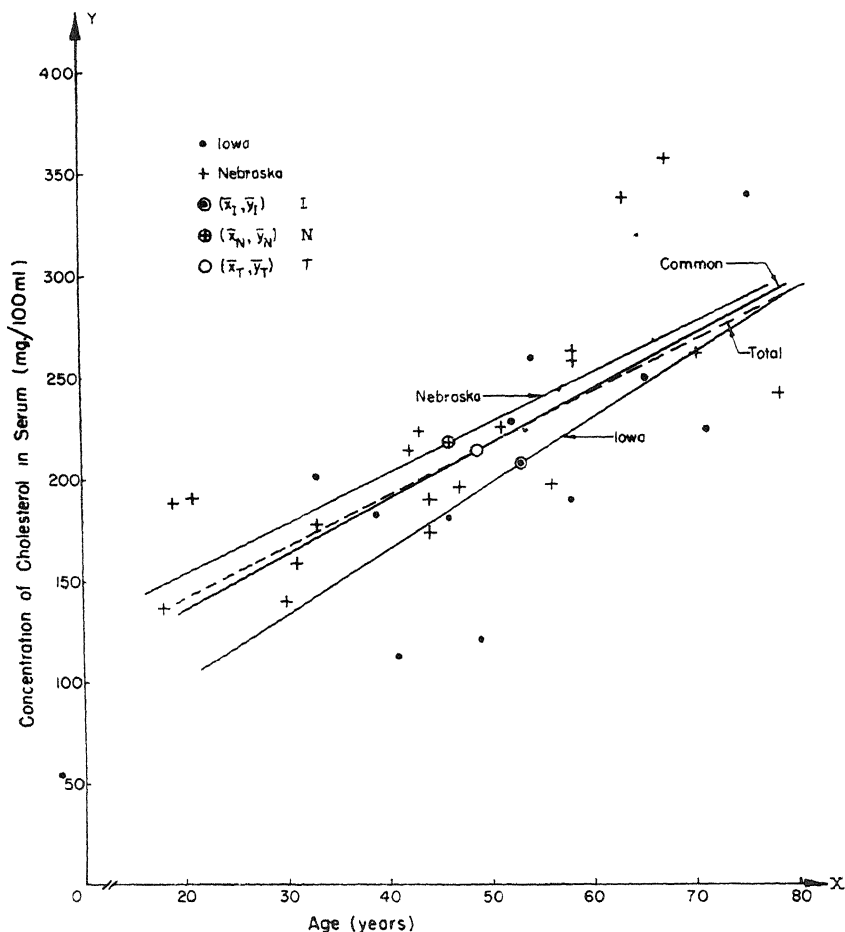


FIG. 14.6.1—Graph of 11 pairs of Iowa data and 19 pairs from Nebraska. Age is X and concentration of cholesterol, Y .

$S.S.$ for the deviations from the individual regression, recording these sums in line 3. The mean square, 1,862, is the residual mean square obtained when *separate* regression lines are fitted in each state. Secondly, in line 4 we add the sums of squares and products, obtaining the pooled slope, 2.70, and the $S.S.$, 49,107, representing deviations from a model in which a *single* pooled slope is fitted. The difference, $49,107 - 48,399 = 708$ (line 5), with 1 $df.$, measures the contribution of the difference between the two regression coefficients to the sum of squares of deviations. If there were k coefficients, this difference would have $(k - 1) df.$ The corresponding mean square is compared with the Within States mean square

TABLE 14.6.2
COMPARISON OF REGRESSION LINES CHOLESTEROL DATA

		<i>df.</i>	Σx^2	Σxy	Σy^2	Reg. Coef	Deviations From Regression		
							<i>df</i>	<i>S.S.</i>	<i>M.S.</i>
L	Within								
i	Iowa	10	1,829	5,922	40,698	3.24	9	21,524	2,392
2	Nebraska	18	5,565	14,026	62,226	2.52	17	26,875	1,581
3							26	48,399	1,862
4	Pooled, <i>W</i>	28	7,394	19,948	102,924	2.70	27	49,107	1,819
5			Difference between slopes				1	708	708
6	Between, <i>B</i>	1	355	-466	613				
7	<i>W</i> + <i>B</i>	29	7,749	19,482	103,537		28	54,557	
8			Between adjusted means				1	5,450	5,450

Comparison of slopes: $F = 708/1,862 = 0.38$ ($df = 1, 26$) *N.S.*

Comparison of elevations: $F = 5,450/1,819 = 3.00$ ($df = 1, 27$) *N.S.*

1,862, by the *F*-test. In these data, $F = 708/1,862 = 0.38$, $df. = 1, 26$, supporting the assumption that the slopes do not differ.

Algebraically, the difference 708 in the sum of squares may be shown to be $\Sigma_1 \Sigma_2 (b_1 - b_2)^2 / (\Sigma_1 + \Sigma_2)$, where Σ_1, Σ_2 are the values of Σx^2 for the two states. With more than two states, the difference is $\Sigma w_i (l_i - \bar{b})^2$ where $w_i = 1/\Sigma_i$ and \bar{b} is the pooled slope, $\Sigma w_i b_i / \Sigma w_i$. The sum of squares of deviations of the b 's is a weighted sum, because the variances of the b 's, namely $\sigma_{y \cdot x}^2 / \Sigma_i$, depend on the values of Σx^2 .

If the sample regressions were found to differ significantly, this might end the investigation. Interpretation would involve the question: Why? The final question about the elevations of the population regression lines usually has little meaning unless the lines are parallel.

Assuming parallel lines and homogeneous variance, we write the model as

$$Y_{ij} = \alpha_i + \beta X_{ij} + \varepsilon_{ij}$$

where $i = 1, 2$, denotes the state. It remains to test the null hypothesis $\alpha_1 = \alpha_2$. The least squares estimates of α_1 and α_2 are $\hat{\alpha}_1 = \bar{Y}_1 - b\bar{X}_1$ and $\hat{\alpha}_2 = \bar{Y}_2 - b\bar{X}_2$. Hence, the test of this H_0 is identical to the test of the H_0 that the adjusted means of the Y 's are the same in the two states. This is, of course, the *F*-test of the difference between adjusted means that was made in section 14.3. It is made in the usual way in line 4 to 8 in table 14.6.2. Line 4 gives the Pooled Within States sums of squares and products, while line 6 shows the Between States sums of squares and products. In line 7 these are combined, just as we combined Error and

Treatments in section 14.3. A Deviations S.S., 54,557, is obtained from line 7 and the Deviations S.S. in line 4 is subtracted to give 5,450, the S.S. Between adjusted means. We find $F = 3.00$, $d.f.$ 1, 27, P about 0.10. In the original survey the difference was smaller than in these subsamples. The investigators felt justified in combining the two states for further examination of the relation between age and cholesterol.

14.7—Comparison of the “Between Classes” and the “Within Classes” regressions. Continuing the theme of section 14.6, we sometimes need to compare the Between Classes regression and the Within Classes regression in the same study. In physiology or biochemistry, for instance, Y and X are measurements made on patients or laboratory animals. Often, the number of subjects is limited, but several measurements of Y and X have been made on each subject. The Between Subjects regression may be the one of primary interest. The objective of the comparison is to see whether the Within and Between regressions appear to estimate the same quantities. If so, they can be combined to give a better estimate of the Between Subjects relationship.

The simplest model that might apply is as follows:

$$Y_{ij} = \alpha + \beta X_{ij} + \varepsilon_{ij}, \quad (14.7.1)$$

where i denotes the class (subject). In this model the same regression line holds throughout the data. The best combined estimates of α and β are obtained by treating the data as a single sample, estimating α and β from the Total line in the analysis of variance.

Two consequences of this model are important: (1) The Between and Within lines furnish independent estimates of β ; call these b_1 and b , respectively. (2) The residual mean squares s_1^2 and s^2 from the regressions in the Between and Within lines are both unbiased estimates of σ^2 , the variance of the ε_{ij} .

To test whether the same regression holds throughout, we therefore compare b_1 and b and s_1^2 and s^2 . Sometimes, b_1 and b agree well, but s_1^2 is found to be much larger than s^2 . One explanation is that all the Y_{ij} for a subject are affected by an additional component of variation d_i , independent of the ε_{ij} . This model is written

$$Y_{ij} = \alpha + \beta X_{ij} + d_i + \varepsilon_{ij}, \quad (14.7.2)$$

If the subjects are a random sample from some population of subjects, the d_i are usually regarded as a random variable from subject to subject with population mean zero and variance σ_B^2 . Under this model, b_1 and b are still unbiased estimates of β , but with m pairs of observations per subject, s_1^2 is an unbiased estimate of $\sigma_1^2 = (\sigma^2 + m\sigma_B^2)$, while s^2 continues to estimate σ^2 . Since the method of comparing b and b_1 and the best way of combining them depend on whether the component d_i is present, we suggest that s^2 and s_1^2 be compared first by an F -test.

The calculations are illustrated by records from ten female leprosy patients. The data are scores representing the abundance of leprosy bacilli at four sites on the body, the X_{ij} being initial scores and the Y_{ij} scores after 48 weeks of a standard treatment. Thus $m = 4$, $n = 10$. (This example is purely for illustration. This regression would probably not be of interest in itself; further, records from many additional patients were available so that a Between Patients regression could be satisfactorily estimated directly.) Table 14.7.1 shows the initial computations.

TABLE 14.7.1
SCORES FOR LEPROSY BACILLI AT FOUR SITES ON TEN PATIENTS

	<i>d.f.</i>	Σx^2	Σxy	Σy^2	Reg. Coef.
Between patients	9	28.00	26.00	38.23	$b_1 = 0.939$
Within patients	30	26.00	13.00	38.75	$b = 0.500$
Total	39	54.00	39.00	76.98	
	Reduction (Σxy) ² / Σx^2	<i>d.f.</i>	Deviations From Regression <i>S.S.</i> <i>M.S.</i>		
Between patients	24.14	8	14.09	$s_1^2 = 1.761$	
Within patients	6.50	29	32.25	$s^2 = 1.112$	

After performing the usual analysis of sums of squares and products, the reduction in sum of squares due to regression is computed separately for the Between and Within lines (lower half of table 14.7.1). From these, the Deviations *S.S.* and *M.S.* are obtained. The *F* ratio is $s_1^2/s^2 = 1.761/1.112 = 1.58$ with 8 and 29 *df.*, corresponding to a *P* level of about 0.20.

Although *F* falls short of significance, the investigator may decide to assume that σ_1^2 is greater than σ^2 , and thus to retain the model (14.7.2), particularly since the Between Patients mean square is significant for both *Y* and *X* individually. To compare b_1 and b under this model, note that the estimated variances of b_1 and b are s_1^2/Σ_1 and s^2/Σ , where Σ_1 and Σ are the values of Σx^2 for Between Patients and Within Patients, respectively. From table 14.7.1 the ratio of $(b_1 - b)$ to its standard error is therefore

$$t' = \frac{b_1 - b}{\sqrt{\frac{s_1^2}{\Sigma_1} + \frac{s^2}{\Sigma}}} = \frac{0.939 - 0.500}{\sqrt{\frac{1.761}{28.00} + \frac{1.112}{26.00}}} = \frac{0.439}{\sqrt{0.0629 + 0.0428}} = \frac{0.439}{0.325} = 1.35$$

which is clearly non-significant. The quantity t' is not distributed as t , but its significance level, if needed, is found by the approximate method in section 4.14. Since s_1^2 has 8 *df.* and s^2 has 29 *df.*, find the 5% sig-

nificance levels of t for 8 *d.f.* and 29 *d.f.*, namely 2.306 and 2.045. Form a weighted mean of these two values, with weights $s_1^2/\Sigma_1 = 0.0629$ and $s^2/\Sigma = 0.0428$. This mean is 2.20, the required 5% significance level of t' .

It remains to find a combined estimate of β from b_1 and b . In combining two independent estimates that are of unequal precision, a general rule is to weight each estimate inversely as its variance. In this example, as is usually the case in practice, we have only estimates $s_1^2/\Sigma_1 = 0.0629$ and $s^2/\Sigma = 0.0428$ of the variances of b_1 and b . If s_1^2 and s^2 both have at least 8 *d.f.*, weight b_1 and b inversely as their estimated variance (8). The weights are $w_1 = 1/0.0629 = 15.9$, $w = 1/0.0428 = 23.4$, giving

$$\hat{\beta} = \frac{(15.9)(0.939) + (23.4)(0.500)}{39.3} = 0.678$$

If $W = w_1 + w = 39.3$, the standard error of $\hat{\beta}$ may be taken as (8)

$$\frac{1}{\sqrt{W}} \sqrt{\left\{ 1 + \frac{4w_1w}{W^2} \frac{(f_1 + f)}{f_1f} \right\}} = 0.171,$$

where f_1, f are the *d.f.* in s_1^2, s^2 . The second term above is an allowance due to Meier (9) for sampling errors in the weights.

We now show how to complete the analysis if $\sigma_1^2 = \sigma^2$. Form a pooled estimate of σ^2 from s_1^2 and s^2 . This is $\hat{\sigma}^2 = 46.34/37 = 1.252$ with 37 *d.f.* The estimated variance of $(b_1 - b)$ is

$$\frac{\hat{\sigma}^2}{\Sigma_1} + \frac{\hat{\sigma}^2}{\Sigma} = \hat{\sigma}^2 \frac{(\Sigma_1 + \Sigma)}{\Sigma_1 \Sigma} = \frac{(1.252)(54.00)}{(28.00)(26.00)} = 0.0929$$

Hence, $(b_1 - b)$ is tested by the ordinary t -test,

$$t = \frac{0.4386}{\sqrt{0.0929}} = \frac{0.4386}{0.305} = 1.44 \quad (37 \text{ d.f.})$$

The pooled estimate of β is simply the estimate $\Sigma xy/\Sigma x^2$ from the Total line in the analysis of variance. This is $39.00/54.00 = 0.722$, with standard error $\sqrt{\{\hat{\sigma}^2/(\Sigma_1 + \Sigma)\}} = \sqrt{(1.252/54.00)} = 0.152$.

Methods for extending this analysis to multiple regression are presented in (10)

14.8—Multiple covariance. With two or more independent variables there is no change in the theory beyond the addition of extra terms in X . The method is illustrated for a one-way classification by the average daily gains of pigs in table 14.8.1. Presumably these are predicted at least partly by the ages and weights at which the pigs were started in the experiment, which compared four feeds.

This experiment is an example of a technique in experimental design known as *balancing*. The assignment of pigs to the four treatments was

not made by strict randomization. Instead, pigs were allotted so that the means of the four lots agreed closely in both X_1 and X_2 . An indication of the extent of the balancing can be seen by calculating the F -ratios for Treatments/Error from the analyses of variance of X_1 and X_2 , given under table 14.8.1. These F 's are 0.50 for X_1 and 0.47 for X_2 , both well below 1.

The idea is that if X_1 and X_2 are linearly related to Y , this balancing produces a more accurate comparison among the Y means. One complication is that since the variance within treatments is greater than that between treatments for X_1 and X_2 , the same happens to some extent for Y . Consequently, in the analysis of variance of Y the Error mean square is an overestimate and the F -test of Y gives too few significant results. However, if the covariance model holds, the analysis of covariance will give an unbiased estimate of error and a correct F -test for the adjusted means of Y . The situation is interesting in that, with balancing, the reason for using covariance is to obtain a proper estimate of error rather than to adjust the Y means. If perfect balancing were achieved, the adjusted Y means would be the same as the unadjusted means.

The first step is to calculate the six sums of squares and products, shown under table 14.8.1. Next, b_1 and b_2 are estimated from the Error lines, the normal equations being

$$\begin{aligned} 4,548.20b_1 + 2,877.40b_2 &= 5.6230 \\ 2,877.40b_1 + 4,876.90b_2 &= 26.2190 \end{aligned}$$

The c_{ij} inverse multipliers are

$$c_{11} = 0.0003508, \quad c_{12} = -0.0002070, \quad c_{22} = 0.0003272$$

These give

$$b_1 = -0.0034542 \quad b_2 = 0.0074142$$

$$\begin{aligned} \text{Reduction in } S.S. &= (-0.0034542)(5.6230) + (0.0074142)(26.2190) \\ &= 0.1750 \end{aligned}$$

$$\text{Deviations } S.S. = 0.8452 - 0.1750 = 0.6702 \quad (34 d.f.): s^2 = 0.0197$$

The standard errors of b_1 and b_2 are

$$s_{b_1} = \sqrt{(s^2 c_{11})} = 0.00263 \quad : \quad s_{b_2} = \sqrt{(s^2 c_{22})} = 0.00254$$

It follows that b_2 is definitely significant but b_1 is not. In practice, we might drop X_1 (age) at this stage and continue the analysis using the regression of Y on X_2 alone. But for illustration we shall adjust for both variables.

If an F -test of the adjusted means is wanted, make a new calculation of b_1 and b_2 from the Treatments plus Error lines, in this case the Total line. The results are $b_1 = -0.0032903$, $b_2 = 0.0074093$, Deviations $S.S. = 0.8415$ (37 d.f.). The F -test is made in table 14.8.2.

The adjusted Y means are computed as follows. In our notation, $\bar{Y}_i, \bar{X}_{1i},$

TABLE 1481
 INITIAL AGE (X_1), INITIAL WEIGHT (X_2), AND RATE OF GAIN (Y) OF 40 PIGS
 (Four treatments in lots of equal size)

	Treatment 1			Treatment 2		
	Initial Age, X_1	Weight, X_2	Gain, Y	Initial Age, X_1	Weight, X_2	Gain, Y
	(days)	(pounds)	(pounds per day)	(days)	(pounds)	(pounds per day)
	78	61	1.40	78	74	1.61
	90	59	1.79	99	75	1.31
	94	76	1.72	80	64	1.12
	71	50	1.47	75	48	1.35
	99	61	1.26	94	62	1.29
	80	54	1.28	91	42	1.24
	83	57	1.34	75	52	1.29
	75	45	1.55	63	43	1.43
	62	41	1.57	62	50	1.29
	67	40	1.26	67	40	1.26
Sums	799	544	14.64	784	550	13.19
Means	79.9	54.4	1.46	78.4	55.0	1.32

	Treatment 3			Treatment 4		
	Initial Age, X_1	Weight, X_2	Gain, Y	Initial Age, X_1	Weight, X_2	Gain, Y
	(days)	(pounds)	(pounds per day)	(days)	(pounds)	(pounds per day)
	78	80	1.67	77	62	1.40
	83	61	1.41	71	55	1.47
	79	62	1.73	78	62	1.37
	70	47	1.23	70	43	1.15
	85	59	1.49	95	57	1.22
	83	42	1.22	96	51	1.48
	71	47	1.39	71	41	1.31
	66	42	1.39	63	40	1.27
	67	40	1.56	62	45	1.22
	67	40	1.36	67	39	1.36
Sums	749	520	14.45	750	495	13.25
Means	74.9	52.0	1.44	75.0	49.5	1.32

	Sums of Squares and Products			
	df	Σx_1^2	$\Sigma x_1 x_2$	Σx_2^2
Treatments	3	187.70	160.15	189.08
Error	36	4,548.20	2,877.40	4,876.90
Total	39	4,735.90	3,037.55	5,065.98

	df	$\Sigma x_1 y$	$\Sigma x_2 y$	Σy^2
Treatments	3	1.3005	1.3218	0.1776
Error	36	5.6230	26.2190	0.8452
Total	39	6.9235	27.5408	1.0228

TABLE 14 8.2
ANALYSIS OF COVARIANCE OF PIG GAINS. DEVIATIONS FROM REGRESSION

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Total	37	0.8415	
Error	34	0.6702	0.0197
For testing adjusted Treatment means	3	0.1713	0.0571*

$F = 0.0571/0.0197 = 2.90^*, df. = 3 \ 34$

\bar{X}_{2i} , denote the means of Y , X_1 , and X_2 for the i th treatment while \bar{X}_1 and \bar{X}_2 denote the overall means of X_1 and X_2 .

Treatment	1	2	3	4	Multiplier
\bar{Y}_i	1.46	1.32	1.44	1.32	1
$(\bar{X}_{1i} - \bar{X}_1)$	+2.9	+1.4	-2.1	-2.0	0.00345 = $-b_1$
$(\bar{X}_{2i} - \bar{X}_2)$	+1.7	+2.3	-0.7	-3.2	-0.00741 = $-b_2$
\bar{Y}_{adj}	1.46	1.31	1.44	1.34	

Thus, for treatment 4,

$$\begin{aligned}\bar{Y}_{adj.} &= \bar{Y}_4 - b_1(\bar{X}_{14} - \bar{X}_1) - b_2(\bar{X}_{24} - \bar{X}_2) \\ &= 1.32 + 0.00345(-2.0) - 0.00741(-3.2) = 1.34\end{aligned}$$

There is little change from unadjusted to adjusted means because of the balancing.

The estimated variance of the difference between the adjusted means of the i th and j th treatments is

$$s^2 [2/n + c_{11}(\bar{X}_{1i} - \bar{X}_{1j})^2 + 2c_{12}(\bar{X}_{1i} - \bar{X}_{1j})(\bar{X}_{2i} - \bar{X}_{2j}) + c_{22}(\bar{X}_{2i} - \bar{X}_{2j})^2]$$

As with covariance on a single X -variable (section 14.2), an average error variance can be used for comparisons among the adjusted means if there are at least 20 $df.$ for Error. The effective Error mean square per observation is

$$s'^2 = s^2 [1 + c_{11}t_{11} + 2c_{12}t_{12} + c_{22}t_{22}]$$

where t_{11} , t_{22} and t_{12} are the Treatments mean squares and mean product. This equation is the extension of (14.2.3) to two X -variables. In these data

$$\begin{aligned}s'^2 &= 0.0197 [1 + \{(0.3508)(62.6) - 2(0.2070)(53.4) + (0.3272)(63.0)\}/10^3] \\ &= (0.0197)(1.020) = 0.0201.\end{aligned}$$

For instance, to find 95% confidence limits for the difference between the adjusted means of treatments 1 and 2, we have

$$D = 1.46 - 1.31 = 0.15 \text{ pounds per day}$$

$$s_D = \sqrt{(2s'^2/10)} = \sqrt{0.00402} = 0.0634$$

The difference 0.15 pounds between Treatments 1 and 2 is the greatest of the six differences between pairs of treatments. It is the only difference that is significant by the *LSD* test. By the Newman-Keuls test, none of the differences is significant, the required difference for 5% significance between the highest and the lowest means being 0.17 pounds. This is one of those occasional examples in which although *F* is significant (just on the 5% level), none of the individual differences between pairs is clearly significant.

These data also illustrate the point that the regression of *Y* on *X*₁ alone may be quite different from the same regression, *Y* on *X*₁, when another *X* variable is included in the model—even the signs may be opposite. Consider the regression of *Y* on *X*₁ (age) in the pig data. Using Totals, the regression coefficient is

$$b_{Y1} = 6.9235/4,735.90 = 0.00146 \text{ lb./day/day of age}$$

Compare this with $b_{Y1.2} = -0.00329$ calculated on p. 439, also for Total. Why should average daily gain increase with age in the first case and decrease with age in the second?

TABLE 14.8.3
DATA ON 40 PIGS CLASSIFIED BY INITIAL WEIGHT

Initial Weight	Number of Pigs	Initial Age and Average Daily Gain								Mean
39-44	13	62 1.57	63(2)* 1.35	66 1.39	67(5) 1.36	70 1.15	71 1.31	83 1.22	91 1.24	69.5 1.34
45-49	5	62 1.22	70 1.23	71 1.39	75(2) 1.45					70.6 1.35
50-54	5	62 1.29	71 1.47	75 1.29	80 1.28	96 1.48				76.8 1.36
55-59	5	71 1.47	83 1.34	85 1.49	90 1.79	95 1.22				84.8 1.46
60-64	8	77 1.40	78(2) 1.38	79 1.73	80 1.12	83 1.41	94 1.29	99 1.26		83.5 1.27
74-80	4	78(2) 1.64	94 1.72	99 1.31						87.2 1.58
Total	40									77.05 1.388

* Number of pigs of this age

The first regression is an overall effect, ignoring initial weight. In this sample there was a slight tendency for the initially older pigs to gain faster. But among pigs of the same initial weight (initial weight held constant) the older pigs tended to gain more slowly.

These facts may be observed in table 14.8.3. The right-hand column shows that both initial age and rate of gain increase with initial weight; they are positively associated because of their common association with initial weight. But within the rows of the table, where initial weight doesn't change much, there is the opposite tendency. The older pigs tend to gain more slowly. Table 14.8.4 gives the within-weight regressions. In the last line is the Pooled regression, -0.00335 . This average differs only slightly from the average, $b_{y1.2} = -0.00329$, estimating the same effect, the regression of average daily gain on initial age in a population of pigs all having the same initial weight.

TABLE 14.8.4
ANALYSIS OF COVARIANCE IN WEIGHT CLASSES OF PIGS

Weight Class	Degrees of Freedom	Sums of Squares and Products			Regression of Y on X_1
		Σx_1^2	$\Sigma x_1 y$	Σy^2	
39-44	12	831.2308	-6.1885	0.1917	-0.007445
45-49	4	113.2000	2.0860	0.0729	0.018428
50-54	4	634.8000	2.5720	0.0427	0.004052
55-59	4	324.8000	-0.6480	0.1819	-0.001995
60-64	7	486.0000	-3.6700	0.2140	-0.007551
74-80	3	354.7500	-3.3375	0.1015	-0.009408
Pooled	34	2,744.7808	-9.1860	0.8047	-0.003347

14.9—Multiple covariance in a 2-way table. As illustration we select data from an experiment (11, 12) carried out in Britain from 1932 to 1937. The objective was to learn how well the wheat crop could be forecast from measurements on a sample of growing plants. During the growing season a uniform series of measurements were taken at a number of places throughout the country. The data in table 14.9.1 are for three seasons at each of six places and are the means of two standard varieties. In the early stages of the experiment it appeared that most of the available information was contained in two variables, shoot height at the time when ears emerge, X_1 , and plant numbers at tillering, X_2 .

For an initial examination of relationships, the data on Y , X_1 , and X_2 should be free of the place and season effects. Consequently, the regression is calculated from the Error or Places \times Seasons Interactions line. If, however, the regression is to be successful for routine use in predicting yields, it should also predict the differences in yield between seasons. It might even predict the differences in yield between places, though this is too much to expect unless the X -variables can somehow express the

effects of differences in soil types and soil fertilities between stations. Consequently, in data of this type, there is interest in comparing the Between Seasons and Between Places regressions with the Error regression, though we shall not pursue this aspect of the analysis.

TABLE 14.9.1
HEIGHTS OF SHOOTS AT EAR EMERGENCE (X_1), NUMBER OF PLANTS AT TILLERING (X_2),
AND YIELD (Y) OF WHEAT IN GREAT BRITAIN
(X_1 , inches; X_2 , number per foot; Y , cwt. per acre)

Year	Variate	Place						Year Sums
		Seale Hayne	Rothamsted	Newport	Bog-hall	Sprows-ton	Plumpton	
1933	X_1	25.6	25.4	30.8	33.0	28.5	28.0	171.3
	X_2	14.9	13.3	4.6	14.7	12.8	7.5	67.8
	Y	19.0	22.2	35.3	32.8	25.3	35.8	170.4
1934	X_1	25.4	28.3	35.3	32.4	25.9	24.2	171.5
	X_2	7.2	9.5	6.8	9.7	9.2	7.5	49.9
	Y	32.4	32.2	43.7	35.7	28.3	35.2	207.5
1935	X_1	27.9	34.4	32.5	27.5	23.7	32.9	178.9
	X_2	18.6	22.2	10.0	17.6	14.4	7.9	90.7
	Y	26.2	34.7	40.0	29.6	20.6	47.2	198.3
Place Sums	X_1	78.9	88.1	98.6	92.9	78.1	85.1	521.7
	X_2	40.7	45.0	21.4	42.0	36.4	22.9	208.4
	Y	77.6	89.1	119.0	98.1	74.2	118.2	576.2

	d.f.	Σx_1^2	$\Sigma x_1 x_2$	Σx_2^2
Places	5	106.34	— 47.06	171.46
Seasons	2	6.26	26.24	139.41
Error	10	117.93	20.17	74.20
Total	17	230.53	— 0.65	385.07

	d.f.	$\Sigma x_1 y$	$\Sigma x_2 y$	Σy^2
Places	5	190.83	— 257.03	629.22
Seasons	2	8.41	— 22.26	124.42
Error	10	142.01	— 21.46	228.66
Total	17	341.25	— 300.75	982.30

The results obtained from the Error line are: $b_1 = 1.3148$, $b_2 = -0.6466$, $\Sigma \hat{y}^2 = 200.59$, $\Sigma d^2 = 28.07$ (8 d.f.). These statistics, with some from the table, lead to the following information:

1. Freed from season and place effects, height of shoots and number of plants together account for

$$\Sigma \hat{y}^2 / \Sigma y^2 = 200.59 / 228.66 = 88\%$$

of the Error sum of squares for yield.

2. The predictive values of the two independent variables are indicated by the following analysis of Σy^2 :

Source	Degrees of Freedom	Sum of Squares	Mean Square
Regression on X_1 and X_2	2	200.59	
{ Regression on X_1 alone	1	171.01	
{ X_2 after X_1	1	29.58	29.58*
{ Regression on X_2 alone	1	6.21	
{ X_1 after X_2	1	194.38	194.38**
Deviations	8	28.07	3.51

While each X accounts for a significant reduction in Σy^2 , shoot height is the more effective.

3. The Error regression equation is

$$\hat{Y} = 1.393 + 1.3148 X_1 - 0.6466 X_2$$

Substituting each pair of X , the values of \hat{Y} and $Y - \hat{Y}$ are calculated for each place in each season and entered in table 14.9.2.

TABLE 14.9.2
ACTUAL AND ESTIMATED YIELDS OF WHEAT

Place	1933			1934			1935			Sum
	Y	\hat{Y}	$Y - \hat{Y}$	Y	\hat{Y}	$Y - \hat{Y}$	Y	\hat{Y}	$Y - \hat{Y}$	
Seale Hayne	19.0	25.4	-6.4	32.4	30.1	2.3	26.2	26.0	0.2	-3.9
Rothamsted	22.2	26.2	-4.0	32.2	32.5	-0.3	34.7	32.3	2.4	-1.9
Newport	35.3	38.9	-3.6	43.7	43.4	0.3	40.0	37.7	2.3	-1.0
Boghall	32.8	35.3	-2.5	35.7	37.7	-2.0	29.6	26.2	3.4	-1.1
Sprowston	25.3	30.6	-5.3	28.3	29.5	-1.2	20.6	23.2	-2.6	-9.1
Plumpton	35.8	33.4	2.4	35.2	28.4	6.8	47.2	39.5	7.7	16.9
Sums			-19.4			5.9			13.4	-0.1

It seems clear from table 14.9.2 that the regression has not been successful in predicting the differences between seasons. There is a consistent underestimation in 1933, which averaged $19.4/6 = 3.2$ cwt./acre, and an overestimation in 1935. If a test of significance of the difference between the adjusted seasonal yields is needed, the procedure is the same as for the F test of adjusted means in section 14.8. Add the sums of squares and products for Seasons and Error in table 14.9.1. Recalculate the regression from these figures, finding the deviations $S.S.$, 120.01 with 10 $d.f.$ The difference, $120.01 - 28.07$ has 2 $d.f.$, giving a mean square 45.97 for the

differences between adjusted seasonal yields. The value of F is $45.97/3.51 = 13.1^{**}$ with 2 and 8 $d.f.$

REFERENCES

1. R. A. FISHER. *Statistical Methods for Research Workers*. §49.1. Oliver and Boyd, Edinburgh (1941).
2. D. J. FINNEY. *Biometrics Bul.*, 2:53 (1946).
3. G. F. SPRAGUE. Iowa Agric. Exp. Sta. data (1952).
4. H. F. SMITH. *Biometrics*, 13:282 (1957).
5. J. M. CRALL. Iowa Agric. Exp. Sta. data (1949).
6. U.S. Bureau of the Census. *Statistical Abstract of the U.S.*, 86th ed. U.S. GPO, Washington, D.C. (1965).
7. P. P. SWANSON *et al.* *J. Gen. mtology*, 10:41 (1955).
8. W. G. COCHRAN. *Biometrics*, 10:116 (1954).
9. P. MEIER. *Biometrics*, 9:59 (1953).
10. D. B. DUNCAN and M. WALSER. *Biometrics*, 22:26 (1966).
11. M. M. BARNARD. *J. Agric. Sci.*, 26:456 (1936).
12. F. YATES. *J. Ministry of Agric.*, 43:156 (1936).
13. G. M. SMITH and H. T. BIECHER. *J. Pharm. and Exner. Therap.*, 136:47 (1962).

Curvilinear regression

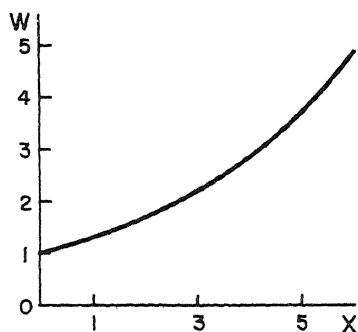
15.1—Introduction. Although linear regression is adequate for many needs, some variables are not connected by so simple a relation. The discovery of a precise description of the relation between two or more quantities is one of the problems of *curve fitting*, known as *curvilinear regression*. From this general view the fitting of the straight line is a special case, the simplest and indeed the most useful.

The motives for fitting curves to non-linear data are various. Sometimes a good estimate of the dependent variable is wanted for any particular value of the independent. This may involve the smoothing of irregular data and the interpolation of estimated Y 's for values of X not contained in the observed series. Sometimes the objective is to test a law relating the variables, such as a growth curve that has been proposed from previous research or from mathematical analysis of the mechanism by which the variables are connected. At other times the form of the relationship is of little interest; the end in view is merely the elimination of inaccuracies which non-linearity of regression may introduce into a correlation coefficient or an experimental error.

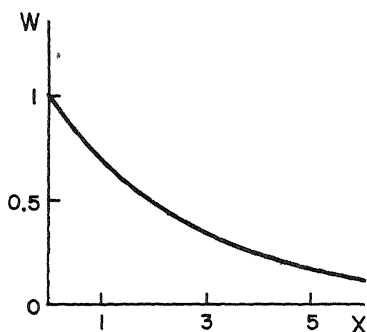
Figure 15.1.1 shows four common non-linear relations. Part (a) is the compound interest law or *exponential growth curve* $W = A(B^X)$, where we have written W in place of our usual Y . If $B = 1 + i$, where i is the annual rate of interest, W gives the amount to which a sum of money A will rise if left at compound interest for X years. As we shall see, this curve also represents the way in which some organisms grow at certain stages. The curve shown in Part (a) has $A = 1$.

If B is less than 1, this curve assumes the form shown in (b). It is often called an *exponential decay curve*, the value of W declining to zero from its initial value A as X increases. The decay of emissions from a radioactive element follows this curve.

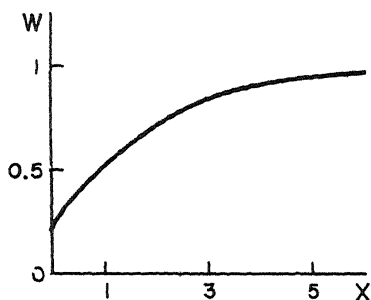
The curve in (c) is $W = A - B\rho^X$, with $0 < \rho < 1$. This curve rises from the value $(A - B)$ when $X = 0$, and steadily approaches a maximum value A , called the asymptote, as X becomes large. The curve goes by various names. In agriculture it has been known as *Mitscherlich's law*, from a German chemist (11) who used it to represent the relation between



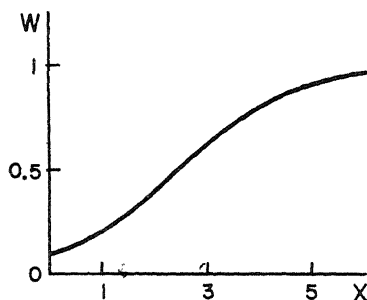
(a) Exponential Growth Law
 $W = A(B^X) = A(e^{cX})$



(b) Exponential Decay Law
 $W = A(B^{-X}) = A(e^{-cX})$



(c) Asymptotic Regression
 $W = A - B(\rho^X) = A - B(e^{-cX})$



(d) Logistic Growth Law
 $W = A/(1 + B\rho^X)$

FIG. 15.1.1—Four common non-linear curves.

the yield W of a crop (grown in pots) and the amount of fertilizer X added to the soil in the pots. In chemistry it is sometimes called the *first-order reaction curve*. The name *asymptotic regression* is also used.

Curve (d), the *logistic growth law*, has played a prominent part in the study of human populations. This curve gives a remarkably good fit to the growth of the U.S. population, as measured in the decennial censuses, from 1790 to 1940.

In this chapter we shall illustrate the fitting of three types of curve: (1) certain non-linear curves, like those in (a) and (b), figure 15.1.1, which can be reduced to straight lines by a transformation of the W or the X scale; (2) the polynomial in X , which often serves as a good approximation; (3) non-linear curves, like (c) and (d), figure 15.1.1, requiring more complex methods of fitting.

EXAMPLE 15.1.1—The fit of the logistic curve of the U.S. Census populations (excluding Hawaii and Alaska) for the 150-year period from 1790 to 1940 is an interesting

example, both of the striking accuracy of the fit, and of its equally striking failure when extrapolated to give population forecasts for 1950 and 1960. The curve, fitted by Pearl and Reed (1), is

$$W = \frac{184.00}{1 + (66.69)(10^{-0.1398X})}$$

where $X = 1$ in 1790, and one unit in X represents 10 years, so that $X = 16$ in 1940. The table below shows the actual census population, the estimated population from the logistic, and the error of estimation.

Year	Population			Year	Population		
	Actual	Estimated	A-E		Actual	Estimated	A-E
1790	3.9	3.7	+0.2	1880	50.2	50.2	0.0
1800	5.3	5.1	+0.2	1890	62.9	62.8	+0.1
1810	7.2	7.0	+0.2	1900	76.0	76.7	-0.7
1820	9.6	9.5	+0.1	1910	92.0	91.4	+0.6
1830	12.9	12.8	+0.1	1920	105.7	106.1	-0.4
1840	17.1	17.3	-0.2	1930	122.8	120.1	+2.7
1850	23.2	23.0	+0.2	1940	131.4	132.8	-1.4
1860	31.4	30.3	+1.1	1950	150.7	143.8	+6.9
1870	38.6	39.3	-0.7	1960	178.5	153.0	+25.5

Note how poor the 1950 and 1960 forecasts are. The forecast from the curve is that the U.S. population will never exceed 184 million; the actual 1966 population is already well over 190 million. The postwar baby boom and improved health services are two of the responsible factors.

15.2—The exponential growth curve. A characteristic of some of the simpler growth phenomena is that the increase at any moment is proportional to the size already attained. During one phase in the growth of a culture of bacteria, the numbers of organisms follow such a law. The relation is nicely illustrated by the dry weights of chick embryos at ages 6 to 16 days (2) recorded in table 15.2.1. The graph of the weights in figure 15.2.1 ascends with greater rapidity as age increases, the regression equation being of the form

$$W = (A)(B^X),$$

where A and B are constants to be estimated. Applying logarithms to the equation,

$$\begin{aligned}\log W &= \log A + (\log B)X \\ \text{or } Y &= \alpha + \beta X,\end{aligned}$$

where $Y = \log W$, $\alpha = \log A$, and $\beta = \log B$. This means that if $\log W$ instead of W is plotted against X , the graph will be linear. By the device of using the logarithm instead of the quantity itself, the data are said to be *rectified*.

The values of $Y = \log W$ are set out in the last column of the table and are plotted opposite X in the figure. The regression equation, com-

TABLE 15.2.1
DRY WEIGHTS OF CHICK EMBRYOS FROM AGES 6 TO 16 DAYS,
TOGETHER WITH COMMON LOGARITHMS

Ages in Days X	Dry Weight, W (grams)	Common Logarithm of Weight Y
6	0.029	-1.538*
7	0.052	-1.284
8	0.079	-1.102
9	0.125	-0.903
10	0.181	-0.742
11	0.261	-0.583
12	0.425	-0.372
13	0.738	-0.132
14	1.130	0.053
15	1.882	0.275
16	2.812	0.449

* From the table of logarithms, one reads $\log 0.029 = \log 2.9 - \log 100 = 0.462 - 2 = -1.538$.

puted in the familiar manner from the columns X and Y in the table, is

$$\bullet \hat{Y} = 0.1959X - 2.689$$

The regression line fits the data points with unusual fidelity, the correlation between Y and X being 0.9992. The conclusion is that the chick embryos, as measured by dry weight, are growing in accord with the exponential law, the logarithm of the dry weight increasing at the estimated uniform rate of 0.1959 per day.

Often, the objective is to learn whether the data follow the exponential law. The graph of $\log W$ against X helps in making an initial judgment on this question, and may be sufficient to settle the point. If so, the use of *semi-logarithmic* graph paper avoids the necessity for looking up the logarithms of W . The horizontal rulings on this graph paper are drawn to such a scale that the plotting of the original data results in a straight line if the data follow the exponential growth law. Semi-log paper can be purchased at most stationery shops. If you require a more thorough method of testing whether the relation between $\log W$ and X is linear, see the end of section 15.3.

For those who know some calculus, the law that the rate of increase at any stage is proportional to the size already attained is described mathematically by the equation

$$\frac{dW}{dX} = cW,$$

where c is the constant *relative* rate of increase. This equation leads to the

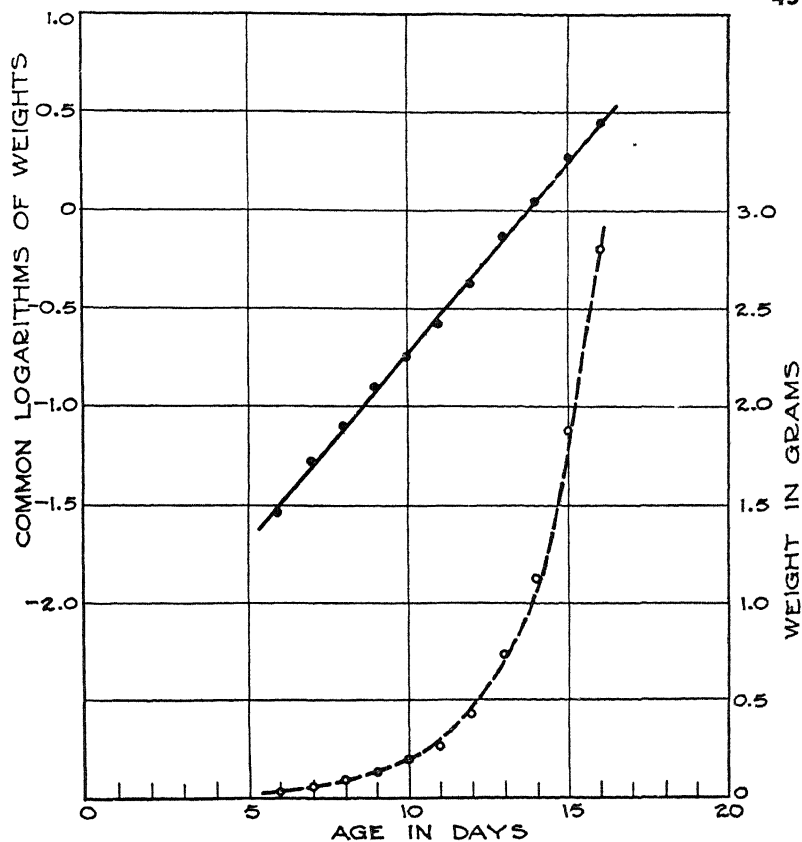


FIG. 15.2.1—Dry weights of chick embryos at ages 6–16 days with fitted curves

Uniform scale: $W = 0.002046(1.57)^X$
 Logarithmic scale: $Y = 0.1959X - 2.689$

relation

$$\log_e W = \log_e A + cX,$$

or,

$$W = Ae^{cX},$$

(15.2.1)

where $e = 2.718$ is the base of the natural system of logarithms. Relation 15.2.1 is exactly the same as our previous relation

$$\log_{10} W = \alpha + \beta X$$

except that it is expressed in logs to base e instead of to base 10.

Since $\log_e W = (\log_{10} W)(\log_e 10) = 2.3026 \log_{10} W$, it follows that $c = 2.3026\beta$. For the chick embryos, the relative rate of growth is

$(2.3026)(0.1959) = 0.451$ gm. per day per gm. It is clear that the relative rate of growth can be computed from either common or natural logs.

To convert the equation $\log \hat{W} = 0.1959X - 2.689$ into the original form, we have

$$\hat{W} = (0.00205)(1.57)^X$$

where $0.00205 = \text{antilog}(-2.689) = \text{antilog}(0.311 - 3) = 2.05/1,000 = 0.00205$. Similarly, $1.57 = \text{antilog}(0.1959)$. In the exponential form,

$$\hat{W} = (0.00205)e^{0.451X}$$

the exponent 0.451 being the relative rate.

Other relations that may be fitted by a simple transformation of the W or the X variable are $W = 1/X$, $W = \alpha + \beta \log X$, and $\log W = \alpha + \beta \log X$. The applicability of the proposed law should first be examined graphically. Should the data appear to lie on a straight line in the relevant transformed scale, proceed with the regression computation. For the last of the above relations, logarithmic paper is available, both vertical and horizontal rulings being in the logarithmic scale.

The transformation of a non-linear relation so that it becomes a straight line is a simple method of fitting, but it involves some assumptions that should be noted. For the exponential growth curve, we are assuming that the population relation is of the form

$$Y = \log W = \alpha + \beta X + \varepsilon, \quad (15.2.2)$$

where the residuals ε are independent, and have zero means and constant variance. Further, if we apply the usual tests of significance to α and β , this involves the assumption that the ε 's are normally distributed. Sometimes it seems more realistic, from our knowledge of the nature of the process or of the measurements, to assume that residuals are normal and have constant variance *in the original W scale*. This means that we postulate a population relation

$$W = (A)(B^X) + d \quad (15.2.3)$$

where A , B now stand for population parameters, and the residuals d are $\mathcal{N}(0, \sigma^2)$.

If equation 15.2.3 holds, it may be shown that in equation 15.2.2 the ε 's will not be normal, and their variances will change as X changes. Given model 15.2.3, the efficient method of fitting is to estimate A and B by minimizing

$$\Sigma (W - AB^X)^2$$

taken over the sample values. This produces non-linear equations in A and B that must be solved by successive approximations. A general method of fitting such equations is given in section 15.7.

EXAMPLE 15.2.1—J. W. Gowen and W. C. Price counted the number of lesions of *Aucuba mosaic virus* developing after exposure to X-rays for various times (data made available through courtesy of the investigators).

Minutes exposure	0	3	7.5	15	30	45	60
Count in hundreds	271	226	209	108	59	29	12

Plot the count as ordinate, then plot its logarithm. Derive the regression, $Y = 2.432 - 0.02227X$, where Y is the logarithm of the count and X is minutes exposure.

EXAMPLE 15.2.2—Repeat the fitting of the last example using natural logarithms. Verify the fact that the rate of decrease of hundreds of lesions per minute per hundred is $(2.3026)(0.02227) = 0.05128$.

EXAMPLE 15.2.3—If the meaning of relative rate isn't quite clear, try this approximate method of computing it. The increase in weight of the chick embryo during the thirteenth day is $1.130 - 0.738 = 0.392$ gram; that is, the average rate during this period is 0.392 gm. per day. But the average weight during the same period is $(1.130 + 0.738)/2 = 0.934$ gm. The relative rate, or rate of increase of each gram, is therefore $0.392/0.934 = 0.42$ gm. per day per gm. This differs from the average obtained in the whole period from 6 to 16 days, 0.451, partly because the average weight as well as the increase in weight in the thirteenth day suffered some sampling variation, and partly because the correct relative rate is based on weight and increase in weight at any instant of time, not on day averages.

15.3—The second degree polynomial. Faced by non-linear regression, one often has no knowledge of a theoretical equation to use. In many instances the second degree polynomial,

$$\hat{Y} = a + bX + cX^2,$$

will be found to fit the data satisfactorily. The graph is a parabola whose axis is vertical, but usually only small segments of such a parabola appear in the process of fitting. Instead of rectifying the data a third variate is added, the square of X . This introduces the methods of multiple regression. The calculations proceed exactly as in chapter 13, X and X^2 being the two independent variates. It need only be remarked that \sqrt{X} , $\log X$, or $1/X$ might have been added instead of X^2 if the data had required it.

To illustrate the method and some of its applications, we present the data on wheat yield and protein content (3) in table 15.3.1 and figure 15.3.1. The investigator wished to estimate the protein content for various yields. We shall also test the significance of the departure from linearity.

The second column of the table contains the squares of the yields in column 1. The squares are treated in all respects like a third variable in multiple regression. The regression equation, calculated as usual,

$$\hat{Y} = 17.703 - 0.3415X + 0.004075X^2,$$

is plotted in the figure. At small values of yield the second degree term with its small coefficient is scarcely noticeable, the graph falling away almost like a straight line. Toward the right, however, the term in X^2 has bent the curve to practically a horizontal direction.

TABLE 15.3 1
PERCENTAGE PROTEIN CONTENT (Y) AND YIELD (X) OF WHEAT
FROM 91 PLOTS*

Yield, Bushel Per Acre X	Square X^2	Percentage Protein Y	Yield, Bushel Per Acre X	Square X^2	Percentage Protein Y
43	1,849	10.7	19	361	13.9
42	1,764	10.8	19	361	13.2
39	1,521	10.8	19	361	13.8
39	1,521	10.2	18	324	10.6
38	1,444	10.3	18	324	13.0
38	1,444	9.8	18	324	13.4
37	1,369	10.1	18	324	13.7
37	1,369	10.4	18	324	13.0
36	1,296	10.3	17	289	13.4
36	1,296	11.0	17	289	13.5
36	1,296	12.2	17	289	10.8
35	1,225	10.9	17	289	12.5
35	1,225	12.1	17	289	12.7
34	1,156	10.4	17	289	13.0
34	1,156	10.8	17	289	13.8
34	1,156	10.9	16	256	14.3
34	1,156	12.6	16	256	13.6
33	1,089	10.2	16	256	12.3
32	1,024	11.8	16	256	13.0
32	1,024	10.3	16	256	13.7
32	1,024	10.4	15	225	13.3
31	961	12.3	15	225	12.9
31	961	9.6	14	196	14.2
31	961	11.9	14	196	13.2
31	961	11.4	12	144	15.5
30	900	9.8	12	144	13.1
30	900	10.7	12	144	16.3
29	841	10.3	11	121	13.7
28	784	9.8	11	121	18.3
27	729	13.1	11	121	14.7
26	676	11.0	11	121	13.8
26	676	11.0	11	121	14.8
25	625	12.8	10	100	15.6
25	625	11.8	10	100	14.6
24	576	9.9	9	81	14.0
24	576	11.6	9	81	16.2
24	576	11.8	9	81	15.8
24	576	12.3	8	64	15.5
22	484	11.3	8	64	14.2
22	484	10.4	8	64	13.5
22	484	12.6	7	49	13.8
21	441	13.0	7	49	14.7
21	441	14.7	6	36	17.2
21	441	11.5	5	25	17.2
21	441	11.0			
20	400	12.8			
20	400	13.0			

* Read from published graph. This accounts for the slight discrepancy between the correlation we got and that reported by the author.

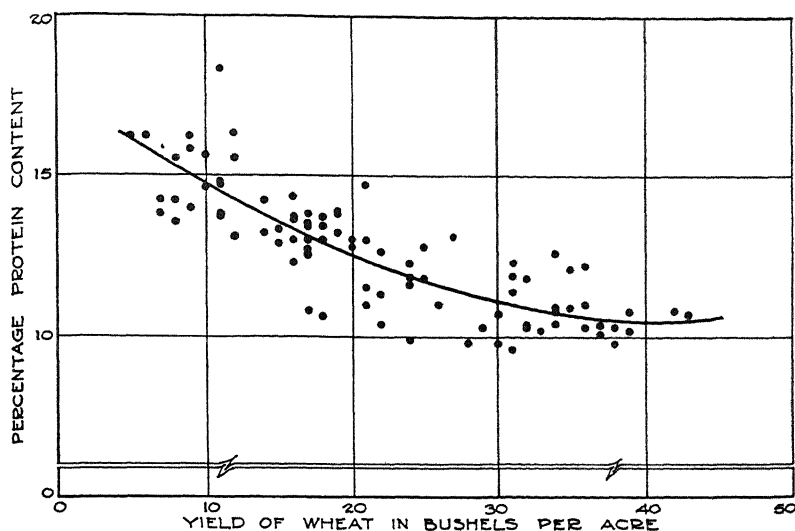


FIG. 15.3.1—Regression of protein content on yield in wheat, 91 plots.
 $Y = 17.703 - 0.3415X + 0.004075X^2$

The analysis of variance and test of significance are shown in table 15.3.2. The fitted regression on both X and X^2 gives a sum of squares of deviations, 97.53, with 88 *df*. The sum of squares of deviations from a linear regression, $\Sigma y^2 - (\Sigma xy)^2 / \Sigma x^2$, is 110.48, with 89 *df*. The reduction in sum of squares, tested against the mean square remaining after curvilinear regression, proves to be significant. The hypothesis of linear regression is abandoned; there is a significant curvilinearity in the regression.

In table 15.4.1, many of the values of X (e.g., $X = 39$) have two or more values of Y . With such data, the sum of squares of deviations from the curved regression (88 *df*.) can be divided into two parts so as to provide a more critical test of the fit of the quadratic. The technique is described in the following section. In the present example this technique supports the quadratic fit.

TABLE 15.3.2
 TEST OF SIGNIFICANCE OF DEPARTURE FROM LINEAR REGRESSION

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Deviations from linear regression	89	110.48	
Deviations from curved regression	88	97.53	1.11
Reduction in sum of squares	1	12.95	12.95**

$$F = 12.95/1.11 = 11.7$$

The regression equation is useful also for estimating and interpolating. Confidence statements and tests of hypotheses are made as in chapter 13.

As always in regression, either linear or curved, one should be wary of extrapolation. The data may be incompetent to furnish evidence of trend beyond their own range. Looking at figure 15.2.1, one might be tempted by the excellent fit to assume the same growth rate before the sixth day and after the sixteenth. The fact is, however, that there were rather sharp breaks in the rate of growth at both these days. To be useful, extrapolation requires extensive knowledge and keen thinking.

EXAMPLE 15.3.1—The test of significance of departure from linear regression in table 15.3.2 may also be used to examine whether a rectifying transformation, of the type illustrated in section 15.2, has produced a straight line relationship. Apply this test to the chick embryo data in table 15.2.1 by fitting a parabola in X to log weights Y . Verify that the parabola is

$$Y = -2.783162 + 0.214503X - 0.000846X^2,$$

and that the test works out as follows:

	Degrees of Freedom	Sum of Squares	Mean Square
Deviations from linear regression	9	0.007094	
Deviations from quadratic regression	8	0.006480	0.000810
Curvilinearity of regression	1	0.000614	0.000614

$F = 0.76$, with 1 and 8 *d.f.* When the X 's are equally spaced, as in this example, a quicker way of computing the test is given in section 15.6.

15.4—Data having several Y 's at each X value. If several values of Y have been measured at each value X , the adequacy of a fitted polynomial can be tested more thoroughly. Suppose that for each X , a group of n values of Y are available. To illustrate for a linear model, if Y_{ij} denotes the j th member of the i th group, the linear model is

$$Y_{ij} = \alpha + \beta X_i + \varepsilon_{ij}, \quad (15.4.1)$$

where the ε_{ij} follow $\mathcal{N}(0, \sigma^2)$. It follows that the group means, \bar{Y}_i , are related to the X_i by the linear relation

$$\bar{Y}_i = \alpha + \beta X_i + \bar{\varepsilon}_i.$$

(1) By fitting a quadratic regression of the \bar{Y}_i on X_i , the test for curvature in table 15.3.2 can be applied as before. Since it is important in what follows, note that the residuals $\bar{\varepsilon}_i$ have variance σ^2/n , since each $\bar{\varepsilon}_i$ is the mean of n independent residuals from relation 15.4.1.

(2) The new feature is that the deviations of the Y_{ij} from their group means \bar{Y}_i supply an independent estimate of σ^2 . The pooled estimate is

$$s^2 = \sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2 / k(n-1)$$

with $k(n - 1)$ *d.f.* If we multiply the mean squares in analysis (1) by n , in order to make parts (1) and (2) comparable, we have the analysis of variance in table 15.4.1.

TABLE 15.4.1
ANALYSIS OF VARIANCE FOR TESTS OF LINEAR REGRESSION

Source of Variation	Degrees of Freedom	Mean Square
Linear regression of \bar{Y}_i on X_i	1	s_1^2
Quadratic regression of \bar{Y}_i on X_i	1	s_2^2
Deviations of \bar{Y}_i from quadratic	$k - 3$	s_d^2
Pooled within groups	$k(n - 1)$	s^2
Total	$kn - 1$	

The following results are basic to the interpretation of this table. If the population regression is linear, the mean square s_2^2 is an unbiased estimate of σ^2 ; if the population regression is curved, s_2^2 tends to become large. If the population regression is either linear or quadratic, s_d^2 is an unbiased estimate of σ^2 . When will s_d^2 tend to become much larger than σ^2 ? *Either* if the population regression is non-linear but is not adequately represented by a quadratic; for instance, it might be a third degree curve, or one with a periodic feature: *or* if there are sources of variation that are constant within any group but vary from group to group. This could happen if the measurements in different groups were taken at different times or from different hospitals or bushes. The pooled within-group variance s^2 is an unbiased estimate of σ^2 no matter what the shape of the relation between \bar{Y}_i and X_i .

Consequently, first compute the F -ratio, s_d^2/s^2 , with $(k - 3)$ and $k(n - 1)$ *d.f.*. If this is significant, look at the plot of Y against X to see whether a higher degree polynomial or a different type of mathematical relationship is indicated. Examination of the deviations of the \bar{Y}_i from the fitted quadratic for signs of a systematic trend is also helpful. If no systematic trend is found, the most likely explanation is that some extra between-group source of variation has entered the data.

If s_d^2/s^2 is clearly non-significant, form the pooled mean square of s_d^2 and s^2 . Call this s_p^2 , with $(kn - 3)$ *d.f.* Then test $F = s_2^2/s_p^2$, with 1 and $(kn - 3)$ *d.f.*, as a test of curvature of the relation.

The procedure is illustrated by the data in table 15.4.2, made available through the courtesy of B. J. Vos and W. T. Dawson. The point at issue is whether there is a linear relation between the lethal dose of ouabain, injected into cats, and the rate of injection. Four rates were used, each double the preceding.

First, the total sum of squares of the lethal doses 21,744 is analyzed into "between rates," 16,093, and "within rate groups," 5,651. Note that the number of cats n_i differed slightly from group to group.

TABLE 15 4 2
LETHAL DOSE (MINUS 30 UNITS) OF U S STANDARD OUABAIN BY SLOW
INTRAVENOUS INJECTION IN CAT UNTIL THE HEART STOPS

X_i = Rate of Injection in (mg /kg /min)/1 045 75					Total
1	2	4	8		
5	3	34	51		
9	6	34	56		
11	22	38	62		
13	27	40	63		
14	27	46	70		
16	28	58	73		
17	28	60	76		
20	37	60	89		
22	40	65	92		
28	42				
31	50				
31					
$\Sigma Y_i = Y_i$	217	310	435	632	1 594
n_i	12	11	9	9	41
\bar{Y}_i	18 1	28 2	48 3	70 2	
ΣY_{ij}^2	4 727	10 788	22 261	45 940	83 716

The inequality in the n_i must be taken into account in setting up the equations for the regression of \bar{Y}_i on X_i and X_i^2 . Compute

$$\begin{aligned}\Sigma n_i X_i &= 12(1) + 11(2) + 9(4) + 9(8) = 142 \\ \Sigma n_i X_i^2 &= 12(1) + 11(4) + 9(16) + 9(64) = 776 \\ \Sigma n_i X_i^3 &= 12(1) + 11(8) + 9(64) + 9(512) = 5\,284\end{aligned}$$

and similarly $\Sigma n_i X_i^4 = 39,356$. We need also

$$\Sigma n_i X_i \bar{Y}_i = \Sigma X_i Y_i = 1(217) + 2(310) + 4(435) + 8(632) = 7,633$$

and $\Sigma X_i^2 Y_i = 48,865$

Each quantity is then corrected for the mean in the usual way. For example,

$$\begin{aligned}\Sigma n_i (X_i^2 - \bar{X}^2)^2 &= \Sigma n_i X_i^4 - (\Sigma n_i X_i^2)^2 / \Sigma n_i = 39,356 - (776)^2 / 41 \\ &= 24,668.8 \\ \Sigma n_i (X_i - \bar{X})(\bar{Y}_i - \bar{Y}) &= \Sigma X_i Y_i - (\Sigma n_i X_i)(\Sigma Y_i) / \Sigma n_i \\ &= 7\,633 - (142)(1,594) / 41 = 2,112.3\end{aligned}$$

To complete the quantities needed for the normal equations, you may verify that

$$\begin{aligned}\Sigma n_i (X_i - \bar{X})^2 &= 284.2, & \Sigma n_i (X_i - \bar{X})(X_i^2 - \bar{X}^2) &= 2,596.4, \\ \Sigma n_i (X_i^2 - \bar{X}^2)(\bar{Y}_i - \bar{Y}) &= 18,695.6\end{aligned}$$

The normal equations for b_1 and b_2 are

$$\begin{aligned} 284.2b_1 + 2,596.4b_2 &= 2,112.3 \\ 2,596.4b_1 + 24,668.8b_2 &= 18,695.6 \end{aligned}$$

In the usual way, the reduction in sum of squares of Y due to the regression on b_1 and b_2 is found to be 16,082, while for the linear regression, the reduction is 15,700. The final analysis of variance appears in table 15.4.3.

TABLE 15.4.3
TESTS OF DEVIATIONS FROM LINEAR AND QUADRATIC REGRESSION

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Linear regression on X	1	15,700	15,700
Quadratic regression on X	1	382	382
Deviations from quadratic	1	11	11
Pooled within groups	37	5,651	153
Total	40	21,744	

The mean square 11 for the deviations from the quadratic is much lower than the Within-groups mean square, though not unusually so for only 1 df . The pooled average of these two mean squares is 149, with 38 df . For the test of curvature, $F = 382/149 = 2.56$, with 1 and 38 df , lying between the 25% and the 10% level. We conclude that the results are consistent with a linear relation in the population.

EXAMPLE 15.4.1—The following data, selected from Swanson and Smith (4) to provide an example with equal n , show the total nitrogen content Y (grams per 100 cc of plasma) of rat blood plasma at nine ages X (days).

Age of Rat	25	37	50	60	80	100	130	180	360
	0.83	0.98	1.07	1.09	0.97	1.14	1.22	1.20	1.16
	0.77	0.84	1.01	1.03	1.08	1.04	1.07	1.19	1.29
	0.88	0.99	1.06	1.06	1.16	1.00	1.09	1.33	1.25
	0.94	0.87	0.96	1.08	1.11	1.08	1.15	1.21	1.43
	0.89	0.90	0.88	0.94	1.03	0.89	1.14	1.20	1.20
	0.83	0.82	1.01	1.01	1.17	1.03	1.19	1.07	1.06
Total	5.14	5.40	5.99	6.21	6.52	6.18	6.86	7.20	7.39

A plot of the Y totals against X shows that (i) the Y values for $Y = 100$ are abnormally low and require special investigation, (ii) the relation is clearly curved. Omit the data for $X = 100$ and test the deviations from a parabolic regression against the Within groups mean square. Ans. $F = 1.4$

15.5—Test of departure from linear regression in covariance analysis.

As in any other correlation and regression work, it is necessary in covariance to be assured that the regression is linear. It will be recalled that in the standard types of layout, one-way classifications, two-way classifications (randomized blocks) and Latin squares, the regression of Y on X is computed from the Residual or Error line in the analysis of variance. A graphical method of checking on linearity, which is often sufficient, is to plot the residuals of Y from the analysis of variance model against the corresponding residuals of X , looking for signs of curvature.

The numerical method of checking is to add a term in X^2 to the model. Writing $X_1 = X$, $X_2 = X^2$, work out the residual or error sums of squares of Y , X_1 , and X_2 , and the error sums of products of X_1X_2 , YX_1 , and YX_2 , as was illustrated in section 14.8 for a one-way classification. From these data, compute the test of significance of departure from linear regression as in table 15.3.2.

If the regression is found to be curved, the treatment means are adjusted for the parabolic regression. The calculations follow the method given in section 14.8.

15.6—Orthogonal polynomials. If the values of X are equally spaced, the fitting of the polynomial

$$Y = b_0 + b_1X + b_2X^2 + b_3X^3 + \dots$$

is speeded up by the use of tables of orthogonal polynomials. The essential step is to replace $X^i (i = 1, 2, 3 \dots)$ by a polynomial of degree i in X , which we will call X_i . The coefficients in these polynomials are chosen so that

$$\sum X_i = 0 \quad : \quad \sum X_i X_j = 0 \quad (i \neq j)$$

where the sums are over the n values of X in the sample. The different polynomials are *orthogonal* to one another. Explicit formulas for these polynomials are given later in this section.

Instead of calculating the polynomial regression of Y on X in the form above, we calculate it in the form:

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \dots$$

which may be shown to give the same fitted polynomial. On account of the orthogonality of the X_i , we have the results:

$$B_0 = \bar{Y} \quad . \quad B_i = \sum X_i Y / \sum X_i^2 \quad (i = 1, 2, 3 \dots)$$

The values of the X_i and of $\sum X_i^2$ are provided in the tables, making the computation of B_i simple. Further, the reductions in $\sum (Y - \bar{Y})^2$ due to the successive terms in the polynomial are given by

$$(\sum X_1 Y)^2 / (\sum X_1^2); \quad (\sum X_2 Y)^2 / (\sum X_2^2); \quad (\sum X_3 Y)^2 / (\sum X_3^2); \quad \text{and so on.}$$

Thus it is easy to check whether the addition of a higher power in X to the

polynomial produces a marked reduction in the residual sum of squares. As a time-saver, the orthogonal polynomials are most effective when the calculations are done on a desk calculator. With an electronic computer, the routine programs for fitting a multiple regression can be used to fit the equation in its original form. Most programs also provide the reductions in sum of squares due to each successive power.

Tables of the first five polynomials are given in (5) up to $n = 75$, and of the first six in (6) up to $n = 52$. Table A 17 (p. 572) shows these polynomials up to $n = 12$. For illustration, a polynomial will be fitted to the chick embryo data, though, as we saw in section 15.2, these data are more aptly fitted as an exponential growth curve.

Table 15.6.1 shows the weights (Y) and the values of X_1, X_2, X_3, X_4, X_5 for $n = 11$, read from table A 17. To save space, most tables give the X_i values only for the upper half of the values of X . In our sample these are the values from $X = 11$ to $X = 16$. The method of writing down the X_i for the lower half of the sample is seen in table 15.6.1. For the terms of odd degree, X_1, X_3 , and X_5 , the signs are changed in the lower half; for terms of even degree, X_2 and X_4 , the signs remain the same.

TABLE 15.6.1
FITTING A FOURTH DEGREE POLYNOMIAL TO CHICK EMBRYO WEIGHTS

Age X	Dry Wt. Y	X_1	X_2	X_3	X_4	X_5	\hat{Y}_4
(days)	(grams)						
6	0.029	-5	15	-30	6	-3	0.026
7	0.052	-4	6	6	-6	6	0.056
8	0.079	-3	-1	22	-6	1	0.086
9	0.125	-2	-6	23	-1	-4	0.119
10	0.181	-1	-9	14	4	-4	0.171
11	0.261	0	-10	0	6	0	0.265
12	0.425	1	-9	-14	4	4	0.434
13	0.738	2	-6	-23	-1	4	0.718
14	1.130	3	-1	-22	-6	-1	1.169
15	1.882	4	6	-6	-6	-6	1.847
16	2.812	5	15	30	6	3	2.822
ΣX_i^2		110	858	4,290	286	156	
λ_i		1	1	5/6	1/12	1/40	
$\Sigma X_i Y$	7.714	25.858	39.768	31.873	1.315	-0.254	
B_i	0.701273	0.235073	0.046349	0.007430	0.004598		

We shall suppose that the objective is to find the polynomial of lowest degree that seems an adequate fit. Consequently, the reduction in sum of squares will be tested as each successive term is added. At each stage,

calculate

$$\Sigma X_i Y, \quad B_i = \Sigma X_i Y / \Sigma X_i^2$$

(shown under table 15.6.1), and the reduction in sum of squares, $(\Sigma X_i Y)^2 / \Sigma X_i^2$, entered in table 15.6.2. For the linear term, the F -value is $(6.078511)/(0.232177) = 26.2$. The succeeding F values for the quadratic and cubic terms are even larger, 59.9 and 173.4. For the X_4 (quartic) term, F is 10.3, significant at the 5% but not at the 1% level. The 5th degree term, however, has an F less than 1. As a precautionary move, we should check the 6th degree term also, but for this illustration we will stop and conclude that a 4th degree polynomial is a satisfactory fit.

TABLE 15.6.2
REDUCTIONS IN SUM OF SQUARES DUE TO SUCCESSIVE TERMS

Source	Degrees of Freedom	Sum of Squares	Mean Square	F
Total: $\Sigma(Y - \bar{Y})^2$	10	8.168108		
Reduction to linear	1	6.078511		
Deviations from linear	9	2.089597	0.232177	26.2
Reduction to quadratic	1	1.843233		
Deviations from quadratic	8	0.246364	0.030796	59.9
Reduction to cubic	1	0.236803		
Deviations from cubic	7	0.009561	0.001366	173.4
Reduction to quartic	1	0.006046		
Deviations from quartic	6	0.003515	0.000586	10.3
Reduction to quintic	1	0.000414		
Deviations from quintic	5	0.003101	0.000620	0.7

For graphing the polynomial, the estimated values \hat{Y} for each value of X are easily computed from table 15.6.1:

$$\hat{Y} = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + B_4 X_4$$

Note that $B_0 = \bar{Y} = 0.701273$. At $X = 6$,

$$\begin{aligned} \hat{Y} &= 0.701273 - 5(0.235073) + 15(0.046349) - 30(0.007430) \\ &\quad + 6(0.004598) = 0.026, \end{aligned}$$

and so on. Figure 15.6.1 shows the fit by a straight line, obviously poor, the 2nd degree polynomial, considerably better, and the 4th degree polynomial.

To express the polynomial as an equation in the original X variables

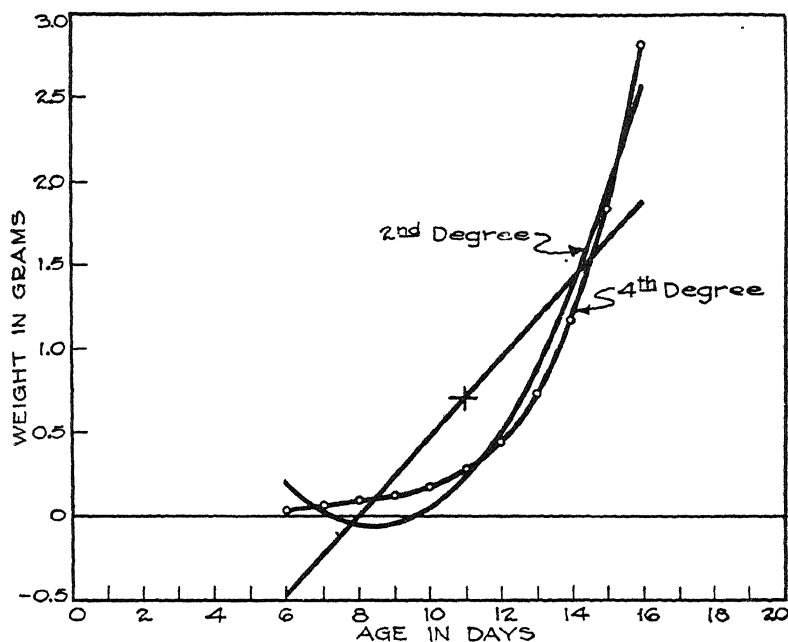


FIG. 15.6.1—Graphs of polynomials of first, second, and fourth degree fitted to chick embryo data of table 15.6.1.

is more tedious. For this, we need formulas giving X_i in terms of X and its powers. In the standard method, developed by Fisher, by which the polynomial tables were computed, he started with a slightly different set of polynomials ξ_i , which satisfy the recurrence relations

$$\xi_0 = 1 : \xi_1 = X - \bar{X} : \xi_{i+1} = \xi_i \xi_1 - \frac{i^2(n^2 - i^2)}{4(4i^2 - 1)} \xi_{i-1}$$

These polynomials are orthogonal, but when their values are tabulated for each member of the sample, these values are not always whole numbers. Consequently, Fisher found by inspection the multiplier λ_i which would make $X_i = \lambda_i \xi_i$ the smallest set of integers. This makes calculations easier for the user. The values of the λ_i are shown under table 15.6.1, and under each polynomial in table A 17 and in references (5) and (6).

Now to the calculations in our example. The first step is to multiply each B_i by the corresponding λ_i . This gives

$$B_1 = 0.235073; \quad B_2' = 0.046349; \quad B_3' = 0.006192; \quad B_4' = 0.0003832$$

These are the coefficients for the regression of Y on the ξ_i , so that

$$\hat{Y} = \bar{Y} + B_1' \xi_1 + B_2' \xi_2 + B_3' \xi_3 + B_4' \xi_4 \quad (15.6.1)$$

The general equations connecting the ξ_i with X are as follows:

$$\xi_1 = X - \bar{X} = x$$

$$\xi_2 = x^2 - \frac{n^2 - 1}{12}$$

$$\xi_3 = x^3 - \frac{3n^2 - 7}{20}x$$

$$\xi_4 = x^4 - \frac{(3n^2 - 13)}{14}x^2 + \frac{3(n^2 - 1)(n^2 - 9)}{560}$$

$$\xi_5 = x^5 - \frac{5(n^2 - 7)}{18}x^3 + \frac{(15n^4 - 230n^2 + 407)}{1,008}x$$

By substitution into formula (15.6.1), \hat{Y} is expressed as a polynomial in $x = X - \bar{X}$. If it is satisfactory to stop at this stage, there are two advantages. Further calculation is avoided, and there is less loss of decimal accuracy. However, to complete the example, we note that $n = 11$ and $\bar{X} = 11$. Hence, in terms of X ,

$$\xi_1 = X - 11$$

$$\xi_2 = (X - 11)^2 - 10 = X^2 - 22X + 111$$

$$\xi_3 = (X - 11)^3 - 17.8(X - 11) = X^3 - 33X^2 + 345.2X - 1,135.2$$

$$\begin{aligned}\xi_4 &= (X - 11)^4 - 25(X - 11)^2 + 72 \\ &= X^4 - 44X^3 + 701X^2 - 4,774X + 11,688\end{aligned}$$

Hence, finally, using formula (15.6.1),

$$\begin{aligned}\hat{Y} &= 0.701273 + 0.235073\xi_1 + 0.046349\xi_2 + 0.006192\xi_3 + 0.0003832\xi_4 \\ &= 0.701273 + 0.235073(X - 11) + 0.046349(X^2 - 22X + 111) \\ &\quad + 0.006192(X^3 - 33X^2 + 345.2X - 1,135.2) \\ &\quad + 0.0003832(X^4 - 44X^3 + 701X^2 - 4,774X + 11,688) \\ &= 0.7099 - 0.47652X + 0.110636X^2 - 0.010669X^3 + 0.0003832X^4\end{aligned}$$

In table 15.6.1 there is a further shortcut which we did not use. In computing $\Sigma X_i Y_i$, the Y 's at the two ends of the sample, say Y_n and Y_1 , are multiplied by 5 and -5 . Y_{n-1} and Y_2 are multiplied by 4 and -4 . If we form the differences, $Y_n - Y_1$, $Y_{n-1} - Y_2$, and so on, only the set of multipliers 5, 4, 3, 2, 1, need be used. This device works for any $\Sigma X_i Y_i$ in which i is odd. With i even, we form the sums $Y_n + Y_1$, $Y_{n-1} + Y_2$, and so on. The method is worked out for these data in example 15.6.1.

EXAMPLE 15.6.1—In table 15.6.1, form the sums and differences of pairs of values of Y , working in from the outside. Verify that these give the results shown below, and that the $\Sigma Y_i Y$ values are in agreement with those given in table 15.6.1.

Sums	X_2	X_4	Diffs.	X_1	X_3
0.261	-10	6	0.261	0	0
0.606	-9	4	0.244	1	-14
0.863	-6	-1	0.613	2	-23
1.209	-1	-6	1.051	3	-22
1.934	6	-6	1.830	4	-6
2.841	15	6	2.783	5	30

EXAMPLE 15.6.2—Here are six points on the cubic, $Y = 9X - 6X^2 + X^3$. (0, 0), (1, 4), (2, 2), (3, 0), (4, 4), (5, 20). Carry through the computations for fitting a linear, quadratic, and cubic regression. Verify that there is no residual sum of squares after fitting the cubic, and that the polynomial values at that stage are exactly the Y 's.

EXAMPLE 15.6.3—The method of constructing orthogonal polynomials can be illustrated by finding X_1 and X_2 when $n = 6$.

(1)	(2)	(3)	(4)	(5)
X	$\xi_1 = X - \bar{X}$	$X_1 = 2\xi_1$	ξ_2	$X_2 = \frac{3}{2}\xi_2$
1	-5/2	-5	10/3	5
2	-3/2	-3	-2/3	-1
3	-1/2	-1	-8/3	-4
4	1/2	1	-8/3	-4
5	3/2	3	-2/3	-1
6	5/2	5	10/3	5

Start with $X = 1, 2, 3, 4, 5, 6$, with $\bar{X} = 7/2$. Verify that the values of $\xi_1 = x = X - \bar{X}$ are as shown in column (2). Since the ξ_1 are not whole numbers, we take $\lambda_1 = 2$, giving $X_1 = 2\xi_1$, column (3). To find ξ_2 , write

$$\xi_2 = \xi_1^2 - b\xi_1 - c$$

This is a quadratic in X . We want $\Sigma \xi_2 = 0$. This gives

$$\Sigma \xi_1^2 - b \Sigma \xi_1 - nc = 0 \quad : \quad \text{i.e., } \frac{35}{2} - 6c = 0 \quad : \quad c = \frac{35}{12}$$

Further, we want $\Sigma \xi_1 \xi_2 = 0$, giving

$$\Sigma \xi_1^3 - b \Sigma \xi_1^2 - c \Sigma \xi_1 = 0 \quad : \quad \text{i.e., } b \Sigma \xi_1^2 = 0 \quad : \quad b = 0$$

Hence, $\xi_2 = \xi_1^2 - \frac{35}{12}$. Verify the ξ_2 values in column (4). To convert these to integers, multiply by $\lambda_2 = \frac{3}{2}$.

15.7—A general method of fitting non-linear regressions. Suppose that the population relation between Y and X is of the form

$$Y_i = f(\alpha, \beta, \gamma, X_i) + \varepsilon_i \quad (i = 1, 2, \dots, n)$$

where f is a regression function containing X_i and the parameters α, β, γ . (There may be more than one X -variable.) If the residuals ε_i have zero means and constant variance, the least squares method of fitting the regres-

sion is to estimate the values of the α , β , γ by minimizing

$$\sum_{i=1}^n [Y_i - f(\alpha, \beta, \gamma, X_i)]^2$$

This section presents a general method of carrying out the calculations. The details require a knowledge of partial differentiation, but the approach is a simple one.

The difficulty arises not because of non-linearity in X_i but because of non-linearity in one or more of the parameters α , β , γ . The parabola ($\alpha + \beta X + \gamma X^2$) is fitted by the ordinary methods of multiple linear regression, because it is linear in α , β , and γ . Consider the asymptotic regression, $\alpha - \beta(\gamma^X)$. If the value of γ were known in advance, we could write $X_1 = \gamma^X$. The least squares estimates of α and β would then be given by fitting an ordinary linear regression of Y on X_1 . When γ must be estimated from the data, however, the methods of linear regression cannot be applied.

The first step in the general method is to obtain good initial estimates a_1 , b_1 , c_1 , of the final least-square estimates $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$. For the common types of non-linear functions, various techniques for doing this have been developed, sometimes graphical, sometimes by special studies of this problem. Next, we use Taylor's theorem. This states that if $f(\alpha, \beta, \gamma, X)$ is continuous in α , β , and γ , and if $(\alpha - a_1)$, $(\beta - b_1)$, and $(\gamma - c_1)$ are small,

$$f(\alpha, \beta, \gamma, X_i) \doteq f(a_1, b_1, c_1, X_i) + (\alpha - a_1)f_a + (\beta - b_1)f_b + (\gamma - c_1)f_c$$

The symbol \doteq means "is approximately equal to." The symbols f_a, f_b, f_c denote the partial derivatives of f with respect to α , β , and γ , respectively, evaluated at the point a_1, b_1, c_1 . For example, in the asymptotic regression,

$$f(\alpha, \beta, \gamma, X_i) = \alpha - \beta(\gamma^{X_i})$$

we have

$$f_a = 1; \quad f_b = -c_1^{X_i}; \quad f_c = -h_1 X_i (c_1^{X_i - 1})$$

Since a_1 , b_1 , and c_1 are known, the values of f , f_a , f_b , and f_c can be calculated for each member of the sample, where we have written f for $f(a_1, b_1, c_1, X_i)$. From Taylor's theorem, the original regression relation

$$Y_i = f(\alpha, \beta, \gamma, X_i) + \varepsilon_i$$

may therefore be written, approximately,

$$Y_i \doteq f + (\alpha - a_1)f_a + (\beta - b_1)f_b + (\gamma - c_1)f_c + \varepsilon_i \quad (15.7.1)$$

Now write

$$Y_{res} = Y - f; \quad X_1 = f_a; \quad X_2 = f_b; \quad X_3 = f_c$$

From equation 15.7.1,

$$Y_{res} \doteq (\alpha - a_1)X_1 + (\beta - b_1)X_2 + (\gamma - c_1)X_3 + \varepsilon_i \quad (15.7.2)$$

The variate Y_{res} is the residual of Y from the first approximation. The relation (15.7.2) represents an ordinary linear regression of Y_{res} on the variates X_1, X_2, X_3 , the regression coefficients being $(\alpha - a_1)$, $(\beta - b_1)$ and $(\gamma - c_1)$. If the relation (15.7.2) held exactly instead of approximately, the computation of the sample regression of Y_{res} on X_1, X_2, X_3 would give the regression coefficients $(\hat{\alpha} - a_1)$, $(\hat{\beta} - b_1)$, and $(\hat{\gamma} - c_1)$, from which the correct least squares estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$ would be obtained at once.

Since relation (15.7.2) is approximate, the fitting of this regression yields second approximations a_2, b_2 , and c_2 to $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$, respectively. We then recalculate f, f_a, f_b and f_c at the point a_2, b_2, c_2 , finding a new Y_{res} and new variates X_1, X_2 , and X_3 . The sample regression of this Y_{res} on X_1, X_2 , and X_3 gives the regression coefficients $(a_3 - a_2), (b_3 - b_2)$ and $(c_3 - c_2)$ from which third approximation a_3, b_3, c_3 to $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ are found, and so on.

If the process is effective, the sum of squares of the residuals, ΣY_{res}^2 , should decrease steadily at each stage, the decreases becoming small as the least-squares solution is approached. In practice, the calculations are stopped when the decrease in ΣY_{res}^2 and the changes in a, b , and c are considered small enough to be negligible. The mean square residual is

$$s^2 = \Sigma Y_{res}^2 / (n - k),$$

where k is the number of parameters that have been estimated (in our example, $k = 3$). With non-linear regression, s^2 is not an unbiased estimate of σ^2 , though it tends to become unbiased as n becomes large.

Approximate standard errors of the estimates $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ are obtained in the usual way from the Gauss multipliers in the final multiple regression that was computed. Thus,

$$s.e. (\hat{\alpha}) \doteq s\sqrt{c_{11}}; \quad s.e. (\hat{\beta}) \doteq s\sqrt{c_{22}}; \quad s.e. (\hat{\gamma}) \doteq s\sqrt{c_{33}}$$

Approximate confidence limits for α are given by $(\hat{\alpha} \pm t s\sqrt{c_{11}})$ where t has $(n - 3) d.f.$

If several stages in the approximation are required, the calculations become tedious on a desk machine, since a multiple regression must be worked out at each stage. With the commonest non-linear relations, however, the computations lend themselves readily to programming on an electronic computer. Investigators with access to a computing center are advised to find out whether a program is available or can be constructed. If the work must be done on a desk machine, the importance of a good first approximation is obvious.

15.8—Fitting an asymptotic regression. The population regression function will be written (using the symbol ρ in place of γ)

$$f(\alpha, \beta, \rho, X) = \alpha + \beta(\rho^X) \quad (15.8.1)$$

If $0 < \rho < 1$ and β is negative, this curve has the form shown in figure 15.1.1(c), p. 448, rising from the value $(\alpha + \beta)$ at $X = 0$ to the asymptote α as X becomes large. If $0 < \rho < 1$ and β is positive, the curve declines from the value $(\alpha + \beta)$ at $X = 0$ to an asymptote α when X is large.

Since the function is non-linear only as regards the parameter ρ , the method of successive approximation described in the preceding section simplifies a little. Let r_1 be a first approximation to ρ . By Taylor's theorem,

$$\alpha + \beta(\rho^X) \doteq \alpha + \beta(r_1^X) + \beta(\rho - r_1)(Xr_1^{X-1})$$

Write $X_0 = 1$, $X_1 = r_1^X$, $X_2 = Xr_1^{X-1}$. If we fit the sample regression

$$\hat{Y} = aX_0 + bX_1 + cX_2 \quad (15.8.2)$$

it follows that a , b are second approximations to the least-squares estimates $\hat{\alpha}$, $\hat{\beta}$, of α and β in (15.8.1), while

$$c = b(r_2 - r_1),$$

so that

$$r_2 = r_1 + c/b \quad (15.8.3)$$

is the second approximation to $\hat{\rho}$.

The commonest case is that in which the values of X change by unity (e.g., $X = 0, 1, 2 \dots$ or $X = 5, 6, 7 \dots$) or can be coded to do so. Denote the corresponding Y values by $Y_0, Y_1, Y_2, \dots, Y_{n-1}$. Note that the value of X corresponding to Y_0 need not be 0. For $n = 4, 5, 6$, and 7, good first approximations to ρ , due to Patterson (7), are as follows:

$$n = 4. \quad r_1 = (4Y_3 + Y_2 - 5Y_1)/(4Y_2 + Y_1 - 5Y_0)$$

$$n = 5. \quad r_1 = (4Y_4 + 3Y_3 - Y_2 - 6Y_1)/(4Y_3 + 3Y_2 - Y_1 - 6Y_0)$$

$$n = 6. \quad r_1 = (4Y_5 + 4Y_4 + 2Y_3 - 3Y_2 - 7Y_1)/(4Y_4 + 4Y_3 + 2Y_2 - 3Y_1 - 7Y_0)$$

$$n = 7. \quad r_1 = (Y_6 + Y_5 + Y_4 - Y_2 - 2Y_1)/(Y_5 + Y_4 + Y_3 - Y_1 - 2Y_0)$$

In a later paper (8), Patterson gives improved first approximations for sample sizes from $n = 4$ to $n = 12$. The value of r_1 , obtained by solving a quadratic equation, is remarkably good in our experience.

In an illustration given by Stevens (9), table 15.8.1 shows six consecutive readings of a thermometer at half-minute intervals after lowering it into a refrigerated hold.

From Patterson's formula (above) for $n = 6$, we find $r_1 = 10.42/-18.86 = 0.552$. Taking $r_1 = 0.55$, compute the sample values of X_1 and X_2 and insert them in table 15.8.1. The matrix of sums of squares and

TABLE 15.8.1
DATA FOR FITTING AN ASYMPTOTIC REGRESSION

X Time (1/2 mins.)	Y Temp. °F	$X_1 =$ (0.55 x)	$X_2 =$ $x(0.55^{x-1})$	\hat{Y}_2	$Y_{res} =$ $Y - \hat{Y}_2$
0	57.5	1.00000	0	57.544	-0.044
1	45.7	0.55000	1.00000	45.525	+0.175
2	38.7	0.30250	1.10000	38.892	-0.193
3	35.3	0.16638	0.90750	35.231	+0.069
4	33.1	0.09151	0.66550	33.211	-0.111
5	32.2	0.05033	0.45753	32.096	+0.104
Total	242.5	2.16072	4.13053		+0.001

products of the three X_i variates is as follows:

$$\begin{array}{lll} \Sigma X_0^2 = 6 & \Sigma X_0 X_1 = 2.16072 & \Sigma X_0 X_2 = 4.13053 \\ \Sigma X_0 X_1 = 2.16072 & \Sigma X_1^2 = 1.43260 & \Sigma X_1 X_2 = 1.11767 \\ \Sigma X_0 X_2 = 4.13053 & \Sigma X_1 X_2 = 1.11767 & \Sigma X_2^2 = 3.68578 \end{array}$$

(Alternatively, we could use the method of sections 13.2-13.4 (p. 381), obtaining a 2×2 matrix of the $\Sigma x_i x_j$, but in the end little time is saved by this.)

The inverse matrix of Gauss multipliers is computed. Each row of this matrix is multiplied in turn by the values of $\Sigma X_i Y$ (placed in the right-hand column).

Inverse matrix			$\Sigma X_i Y$
$c_{11} = 1.62101$	$c_{12} = -1.34608$	$c_{13} = -1.40843$	242.5
$c_{12} = -1.34608$	$c_{22} = 2.03212$	$c_{23} = 0.89229$	104.86457
$c_{13} = -1.40843$	$c_{23} = 0.89229$	$c_{33} = 1.57912$	157.06527

These multiplications give

$$a = 30.723; \quad b = 26.821; \quad c = b(r_2 - r_1) = 0.05024 \quad (15.8.4)$$

Hence,

$$r_2 = r_1 + c/b = 0.55 + 0.05024/26.821 = 0.55187$$

The second approximation to the curve is

$$\hat{Y}_2 = 30.723 + 26.821(0.55187)^x \quad (15.8.5)$$

In order to judge whether the second approximation is near enough to the least-squares solution, we find ΣY_{res}^2 for the first two approximations. The first approximation is

$$\hat{Y}_1 = \hat{a}_1 + b_1(0.55^x) = a_1 + b_1 X_1 \quad (15.8.6)$$

where a_1, b_1 are given by the linear regression of Y on X_1 . In the preceding calculations, a_1 and b_1 were not computed, since they are not needed in finding the second approximation. However, by the usual rules for linear regression, ΣY_{rv}^2 from the first approximation is given by

$$\Sigma Y^2 - (\Sigma Y)^2/n - (\Sigma YX_1)^2/\Sigma X_1^2, \quad (15.8.7)$$

where, as usual, $x_1 = X_1 - \bar{X}_1$. When the curve fits closely, as in this example, ample decimals must be carried in this calculation, as Stevens (9) has warned. Alternatively, we can compute a_1 and b_1 in (15.8.6) and hence $Y - \hat{Y}_1$, obtaining the residual sum of squares directly. With the number of decimals that we carried, we obtained 0.0988 by formula 15.8.7 and 0.0990 by the direct method, the former figure being the more accurate.

For the second approximation, compute the powers of $r_2 = 0.55187$, and hence find \hat{Y}_2 by (15.8.5). The values of \hat{Y}_2 and of $Y - \hat{Y}_2$ are shown in table 15.8.1. The sum of squares of residuals is 0.0973. The decrease from the first approximation (0.0988 to 0.0973) is so small that we may safely stop with the second approximation. Further approximations lead to a minimum of 0.0972.

The Residual mean square for the second approximation is $s^2 = 0.0973/3 = 0.0324$, with $n - 3 = 3$ d.f. Approximate standard errors for the estimated parameters are (using the inverse matrix):

$$\begin{aligned} s.e.(a_2) &= s\sqrt{c_{11}} = \pm 0.23; & s.e.(b_2) &= s\sqrt{c_{22}} = \pm 0.26; \\ s.e.(r_2) &= s\sqrt{c_{33}/h_2} = 0.226/26.82 = \pm 0.0084 \end{aligned}$$

Strictly speaking, the values of the c_{ii} should be calculated for $r = 0.55187$ instead of $r = 0.55$, but the above results are close enough. Further, since $r_2 - r_1 = c/b$, a better approximation to the standard error of r_2 is given by the formula for the standard error of a ratio.

$$s.e.(r_2) = \frac{sc}{b} \left\{ \frac{c_{33}}{c^2} + \frac{c_{22}}{b^2} - \frac{2c_{23}}{bc} \right\}^{\frac{1}{2}}$$

In nearly all cases, the term c_{33}/c^2 in the square root dominates, reducing the result to $s\sqrt{c_{33}/b}$.

When Y has the values 0, 1, 2, . . . , $(n - 1)$, desk machine calculation of the second approximation is much shortened by auxiliary tables. The c_{ii} and c_{ij} in the 3×3 inverse matrix that we must compute at each stage depend only on n and r . Stevens (9) tabulated these values for $n = 5, 6, 7$. With these tables, the user finds the first approximation r_1 , and computes the sample values of X_1 and X_2 and the quantities ΣY , $\Sigma X_1 Y$, $\Sigma X_2 Y$. The values of the c_{ij} corresponding to r_1 are then read from Stevens' tables, and the second approximations are obtained rapidly as in (15.8.4) above. Hiorns (10) has tabulated the inverse matrix for r going by 0.01 from 0.1 to 0.9 and for sample sizes from 5 to 50.

EXAMPLE 15.8.1—In an experiment on wheat in Australia, fertilizers were applied at a series of levels with these resulting yields.

Level	X	0	10	20	30	40
Yield	Y	26.2	30.4	36.3	37.8	38.6

Fit a Mitscherlich equation. Ans. Patterson's formula gives $r_1 = 0.40$. The second approximation is $r_2 = 0.40026$, but the residual sum of squares is practically the same as for the first approximation, which is $\hat{Y} = 38.679 - 12.425(0.4)^x$.

EXAMPLE 15.8.2—In a chemical reaction, the amount of nitrogen pentoxide decomposed at various times after the start of the reaction was as follows (12).

Time (T)	2	3	4	5	6	7
Amount Decomposed (Y)	18.6	22.6	25.1	27.2	29.1	30.1

Fit an asymptotic regression. We obtained $\hat{Y} = 33.802 - 26.698(0.753)^T$, with residual $S.S. = 0.105$.

EXAMPLE 15.8.3—Stevens (9) has remarked that when ρ is between 0.7 and 1, the asymptotic regression curve is closely approximated by a second degree polynomial. The asymptotic equation $Y = 1 - 0.9(0.8)^x$ takes the following values:

X	0	1	2	3	4	5	6
Y	0.100	0.280	0.424	0.539	0.631	0.705	0.764

Fit a parabola by orthogonal polynomials and observe how well the values of Y agree.

REFERENCES

1. R. PEARL, L. J. REED, and J. F. KISH. *Science*, 92:486, Nov. 22 (1940).
2. R. PENQUITE. Thesis submitted for the Ph.D. degree, Iowa State College (1936).
3. W. H. METZGER. *J. Amer. Soc. Agron.*, 27:653 (1935).
4. P. P. SWANSON and A. H. SMITH. *J. Biol. Chem.*, 97:745 (1932).
5. R. A. FISHER and F. YATES. *Statistical Tables*. Oliver and Boyd, Edinburgh, 5th ed. (1957).
6. E. S. PEARSON and H. O. HARTLEY. *Biometrika Tables for Statisticians*, Vol. I. Cambridge University Press (1954).
7. H. D. PATTERSON. *Biometrics*, 12:323 (1956).
8. H. D. PATTERSON. *Biometrika*, 47:177 (1960).
9. W. L. STEVENS. *Biometrics*, 7:247 (1951).
10. R. W. HIORNS. *The Fitting of Growth and Allied Curves of the Asymptotic Regression Type by Stevens's Method*. Tracts for Computers No. XXVIII. Cambridge University Press (1965).
11. E. A. MITSCHERLICH. *Landw. Jahrb.*, 38:537 (1909).
12. L. J. REED and E. J. THERIAULT. *J. Physical Chem.*, 35:950 (1931).

Two-way classifications with unequal numbers and proportions

16.1—Introduction. For one reason or another the numbers of observations in the individual cells (sub-classes) of a multiple classification may be unequal. This is the situation in many non-experimental studies, in which the investigator classifies his sample according to the factors or variables of interest, exercising no control over the way in which the numbers fall. With a one-way classification, the handling of the “unequal numbers” case was discussed in section 10.12. In this chapter we present methods for analyzing a two-way classification. The related problem of analyzing a proportion in a two-way table will be taken up also.

The complications introduced by unequal sub-class numbers can be illustrated by a simple example. Two diets were compared on samples of 10 rats. As it happened, 8 of the 10 rats on Diet 1 were females, while only 2 of the 10 rats on Diet 2 were females. Table 16.1.1 shows the sub-class totals for gains in weight and the sub-class numbers. The 8 females on Diet 1 gained a total of 160 units, and so on.

TABLE 16.1.1
TOTAL GAINS IN WEIGHT AND SUB-CLASS NUMBERS (ARTIFICIAL DATA)

		Females	Males	Sums	Means
Diet 1	Totals	160	60	220	22
	Numbers	8	2	10	
Diet 2	Totals	30	200	230	23
	Numbers	2	8	10	
Sums	Totals	190	260	450	
	Numbers	10	10	20	
Means		19	26		22.5

From these data we obtain the row totals and means, and likewise the column totals and means. From the row means, it looks as if Diet 2 had

a slight advantage over Diet 1, 23 against 22. In the column means, males show greater gains than females, 26 against 19.

The sub-class means per rat tell a different story.

	Female	Male
Diet 1	20	30
Diet 2	15	25

Diet 1 is superior by 5 units in both Females and Males. Further, Males gain 10 units more than Females under both diets, as against the estimate of 7 units obtained from the overall means.

Why do the row and column means give distorted results? Clearly, because of the inequality in the sub-class numbers. The poorer feed, Diet 2, had an excess of the faster-growing males. Similarly, the comparison of Male and Female means is biased because most of the males were on the inferior diet.

If we attempt to compute the analysis of variance by elementary methods, this also runs into difficulty. From table 16.1.1 the sum of squares between sub-classes is correctly computed as

$$\frac{(160)^2}{8} + \frac{(60)^2}{2} + \frac{(30)^2}{2} + \frac{(200)^2}{8} - \frac{(450)^2}{20} = 325 \text{ (3 d.f.)}$$

The sum of squares for Diets, $(230 - 220)^2/20$, is 5, and that for Sex $(260 - 190)^2/20$, is 245, leaving an Interaction sum of squares of 75. But from the cell means there is obviously no interaction; the difference between the Diet means is the same for Males as for Females. In a correct analysis, the Interaction sum of squares should be zero.

For a correct analysis of a two-way table the following approach is suggested:

1. First test for interactions: methods of doing this will be described presently.

2a. If interactions appear negligible, this means that an additive model

$$\bar{X}_{ij} = \mu + \alpha_i + \beta_j + \bar{\epsilon}_{ij}.$$

is a satisfactory fit, where \bar{X}_{ij} is the mean of the n_{ij} observations in the i th row and j th column. Proceed to find the best unbiased estimates of the α_i and β_j .

2b. If interactions are substantial, examine the row effects separately in each column, and vice versa, with a view to understanding the nature of the interactions and writing a summary of the results. The overall row and column effects become of less interest, since the effect of each factor depends on the level of the other factor.

Unfortunately, with unequal cell numbers the exact test of the null hypothesis that interactions are absent requires the solution of a set of

linear equations like those in a multiple regression. Consequently, before presenting the exact test (section 16.7) we first describe some quicker methods that are often adequate. When interactions are large, this fact may be obvious by inspection, or can sometimes be verified by one or two *t*-tests, as illustrated in section 16.2. Also, the exact test can be made by simple methods if the cell numbers n_{ij} are (i) equal, (ii) equal within any row or within any column, or (iii) proportional—that is, in the same proportion within any row. If the actual cell numbers can be approximated reasonably well by one of these cases, an approximate analysis is obtained by using the actual cell means, but replacing the cell numbers n_{ij} by the approximations. The three cases will be illustrated in turn in sections 16.2, 16.3, and 16.4.

The fact that elementary methods of analysis still apply when the cell numbers are proportional is illustrated in table 16.1.2. In this, the cell means are exactly the same as in table 16.1.1, but males and females are now in the ratio 1:3 in each diet, there being 4 males and 12 females on Diet 1 and 1 male and 3 females on Diet 2. Note that the overall row means show a superiority of 5 units for Diet 1, just as the cell means do.

TABLE 16.1.2
EXAMPLE OF PROPORTIONAL SUB-CLASS NUMBERS

	Females		Males		Sums	
	Totals	Numbers	Totals	Numbers	Totals	Numbers
Diet 1	240	12	120	4	360	16
Means		20		30		22.5
Diet 2	45	3	25	1	70	4
Means		15		25		17.5
Sums	285	15	145	5	430	20
Means		19.0		29.0		21.5

Analysis of Variance
Correction term $C = (430)^2/20 = 9,245$

Degrees of Freedom		Sum of Squares	
Rows	1	$\frac{(360)^2}{16} + \frac{(70)^2}{4} - C$	= 80
Columns	1	$\frac{(145)^2}{5} + \frac{(285)^2}{15} - C$	= 375
Interactions	1	By subtraction	= 0
Between sub-classes	3	$\frac{(120)^2}{4} + \frac{(45)^2}{3} - C$	= 455

Similarly, the overall column means show that the males gained 10 units more per animal than females. In the analysis of variance, the Interactions sum of squares is now identically zero.

16.2—Unweighted analysis of cell means. Let X_{ijk} denote the k th observation in the cell that is in the i th row and j th column, while $\bar{X}_{ij\cdot}$ is the cell mean, based on n_{ij} observations. In this method the $\bar{X}_{ij\cdot}$ are treated as if they were all based on the same number of observations when computing the analysis of variance. The only new feature is how to include the Within-cells mean square $s^2 = \Sigma\Sigma\Sigma(X_{ijk} - \bar{X}_{ij\cdot})^2 / \Sigma\Sigma(n_{ij} - 1)$ in the analysis of variance.

With fixed effects, the general model for a two-way classification may be written

$$X_{ijk} = \mu + \alpha_i + \beta_j + I_{ij} + \varepsilon_{ijk}, \quad (16.2.1)$$

where α_i and β_j are the additive row and column effects, respectively. The I_{ij} are population parameters representing the interactions. The I_{ij} sum to zero over any row and over any column, since they measure the extent to which the additive row and column effects fail to fit the data in the body of the two-way table. The ε_{ijk} are independent random residuals or deviations, usually assumed to be normally distributed with zero means and variance σ^2 . It follows from 16.2.1 that for a cell mean,

$$\bar{X}_{ij\cdot} = \mu + \alpha_i + \beta_j + I_{ij} + \bar{\varepsilon}_{ij\cdot},$$

where $\bar{\varepsilon}_{ij\cdot}$ is the mean of n_{ij} deviations.

The variance of $\bar{X}_{ij\cdot}$ is σ^2/n_{ij} . Consequently, if there are a rows and b columns, the average variance of a cell mean is

$$\frac{\sigma^2}{ab} \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \dots + \frac{1}{n_{ab}} \right) = \frac{\sigma^2}{n_h},$$

where n_h is known in mathematics as the *harmonic mean* of the n_{ij} . A table of reciprocals helps in its calculation. The Within-cell mean square is entered in the analysis of variance as s^2/n_h .

Our example (table 16.2.1) comes from an experiment (1) in which 3 strains of mice were inoculated with 3 isolations (i.e., different types) of the mouse typhoid organism. The n_{ij} and the $\bar{X}_{ij\cdot}$ (mean days-to-death) are shown for each cell. The unweighted analysis of variance is given under the table. From the original data, not shown here, s^2 is 5.015 with 774 *d.f.* Since $1/n_h$ was found to be 0.01678, the Within-cells mean square is entered as $(0.01678)(5.015) = 0.0841$ in the analysis of variance table.

The unweighted analysis may be used either as the definitive analysis, or merely as a quick initial test for interactions. As a final analysis the unweighted method is adequate only if the disparity in the n_{ij} is small—say within a 2 to 1 ratio with most cells agreeing more closely. Table 16.2.1

TABLE 16.2.1
CELL NUMBERS AND MEAN DAYS-TO-DEATH IN THREE STRAINS OF MICE INOCULATED
WITH THREE ISOLATIONS OF THE TYPHOID BACILLUS

Isolation	Strain of Mice			Sums
	RI	Z	Ba	
9D n_{ij}	34	31	33	
\bar{X}_{ij}	4.0000	4.0323	3.7576	11.7899
11C	66	78	113	
	6.4545	6.7821	4.3097	17.5463
DSC 1	107	133	188	
	6.6262	7.8045	4.1277	18.5584
Sums	17.0807	18.6189	12.1950	47.8946

Analysis of Variance of Unweighted Means

	Degrees of Freedom	Sum of Squares	Mean Square
Isolations	2	8.8859	
Strains	2	7.5003	
Interactions	4	3.2014	0.8004**
Within cells	774		0.0841

$$\frac{1}{n_h} = \frac{1}{9} \left(\frac{1}{34} + \frac{1}{78} + \frac{1}{188} \right) = 0.01678 \quad n_h = 59.61$$

does not come near to meeting this restriction: the n_{ij} range from 31 to 188. However, this experiment is one in which the presence of interactions would be suspected from a preliminary glance at the data. It looks as if strain Ba was about equally susceptible to all three isolations, while strains RI and Z were more resistant to isolations 11C and DSC1 than to 9D. In this example the unweighted analysis would probably be used only to check this initial impression that an additive model does not apply. The F -ratio for Interactions is $0.8004/0.0841 = 9.51$ with 4 and 774 *d.f.*, significant at the 1% level. Since the additive model is rejected, no comparisons among row and column means seem appropriate.

For subsequent t -tests that are made to aid the interpretation of the results, the method of unweighted means, if applied strictly, regards every cell mean as having an error variance 0.0841. This amounts to assuming that every cell has a sample size $n_h = 59.61$. However, comparisons among cell means can be made without assuming the numbers to be equal. For instance, in examining whether strain Z is more resistant to DSC1 than to 11C, the difference in mean days-to-death is $7.8045 - 6.7821 = 1.0224$, with standard error

$$\sqrt{(5.015) \left(\frac{1}{78} + \frac{1}{133} \right)} = \pm 0.319$$

so that the difference is clearly significant by a t -test. Similarly, in testing whether Ba shows any differences from strain to strain in mean days to death, we have a one-way classification with unequal numbers per class (see example 16.2.1).

If interactions had been negligible, main effects would be estimated approximately from the row and column means of the sub-class means. These means can also be assigned correct standard errors. For instance, for 9D the mean, $11.7899/3 = 3.9300$, has a standard error

$$\sqrt{\frac{(5.015)}{9} \left(\frac{1}{34} + \frac{1}{31} + \frac{1}{33} \right)}$$

In some applications it is suspected that the Within-sub-class variance is not constant from one sub-class to another. Two changes in the approximate method are suggested. In the analysis of variance, compute the Within-classes mean square as the average of the quantities s_{ij}^2/n_{ij} , where s_{ij}^2 is the mean square within the i, j sub-class. In a comparison $\Sigma L_{ij}\bar{X}_{ij}$, among the sub-class means, compute the standard error as

$$\sqrt{\Sigma L_{ij} \bar{s}_{ij}^2 / n_{ij}}$$

using only the sub-classes that enter into the comparison.

EXAMPLE 16.2.1—Test whether Ba shows any differences from strain to strain in mean days-to-death. **Ans** The Ba totals are 124, 487, 776, for sample sizes 33, 113, 188. The weighted sum of squares is 8.0565, with 2 *d.f.* The mean square, 4.028, as compared with the Within-class mean square, 5.015, shows no indication of any difference

16.3—Equal numbers within rows. In the mice example (table 16.2.1), an analysis that assumes equal sub-class numbers within each row approximates the actual numbers much more closely than the assumption that all numbers are equal. Since the row total numbers are 98, 257, and 428, we assign sample sizes 33, 86, and 143 to the sub-classes in the respective rows.

In the analysis (table 16.3.1), each sub-class mean is multiplied by the assigned sub-class number to form a corresponding sub-class total. Thus, for Z with 9D, $133.1 = (33)(4.0323)$. The analysis of variance, given under table 16.3.1, is computed by elementary methods. Each total, when squared, is divided by the assigned sample size.

The F -ratio for Interactions is 8.70, again rejecting the hypothesis of additivity of Isolation and Strain effects. In this example, the assigned numbers agree nearly enough with the actual numbers so that further t -tests may be based on the assigned numbers. If the interactions had been unimportant in this example, the main effects of Isolations and Strains would be satisfactorily estimated from the overall means 3.930, 5.849, and so on, shown in table 16.3.1. (These means were not used in the present calculations.)

TABLE 16.3.1
ANALYSIS OF MICE DATA BY EQUAL NUMBERS WITHIN ROWS
(Assigned numbers \bar{n}_i sub-class means \bar{X}_{ij} , and corresponding totals, $\bar{n}_i \bar{X}_{ij}$.)

Isolation		Strain of Mice			Sums	Means
		RI	Z	Ba		
9D	\bar{n}_i	33	33	33	99	
	\bar{X}_{ij}	4.0000	4.0323	3.7576		
	$\bar{n}_i \bar{X}_{ij}$	132.0	133.1	124.0	389.1	3.930
11C		86	86	86	258	
		6.4545	6.7821	4.3097		
		555.1	583.3	370.6	1,509.0	5.849
DSC1		143	143	143	429	
		6.6262	7.8045	4.1277		
		947.5	1,116.0	590.3	2,653.8	6.186
Sums		262	262	262	786	
		1,634.6	1,832.4	1,084.9	4,551.9	
Means		6.239	6.994	4.141		

Correction: $C = (4,551.9)^2 / 786 = 26,361.060$

Between Sub-classes: $\frac{(132.0)^2}{33} + \frac{(590.3)^2}{143} - C = 1,730.22$

Isolations: $\frac{(389.1)^2}{99} + \frac{(1,509.0)^2}{258} + \frac{(2,653.8)^2}{429} - C = 410.56$

Strains: $\frac{(1,634.6)^2 + (1,832.4)^2 + (1,084.9)^2}{262} - C = 1,145.10$

Analysis of Variance

	Degrees of Freedom	Sum of Squares	Mean Square	F
Isolations	2	410.56	205.28	
Strains	2	1,145.10	572.55	
Interactions	4	174.56	43.64	8.70
Between sub-classes	8	1,730.22		
Within sub-classes	774		5.015	

Although this method requires slightly more calculation than the assumption of equal numbers, it is worth considering if it produces numbers near to the actual numbers.

16.4—Proportional sub-class numbers. As mentioned in section 16.1, the least squares analysis can be carried out by simple methods if the sub-class numbers are in the same proportions within each row. Points to note are:

- (i) The overall row means, found by adding all the observations in a

row and dividing by the sum of the sub-class numbers in the row, are the least squares estimates of the row main effects, and similarly for columns.

(ii) In computing the analysis of variance, the squared total for any sub-class, row, or column is divided by the corresponding number. The Total sum of squares between sub-classes and the sums of squares for rows and columns are calculated directly, the Interaction sum of squares being found by subtraction.

(iii) The F -ratio of the Interactions mean square to the Within sub-classes mean square gives the exact least squares test of the null hypothesis that there are no interactions.

Two examples will be presented. In table 16.4.1 the classes are Breeds of Swine and Sex of Swine. The sub-class numbers represent approximately the proportions in which the breeds and sexes were brought in for slaughter at the College Meats Laboratory (2). For each breed, males and females are in the proportions 2:1, and for each sex, the breeds are in the proportions 6:15:2:3:5. The data are the percentages of dressed weight to total weight (less 70%). The calculations are given in full under the table. Since the sample represents only a small fraction of the original data, conclusions are tentative. There were differences among breeds but no indication of a sex difference nor of sex-breed interactions. In making comparisons among the breed means, account should of course be taken of the differences in the sample sizes.

In the breed means, the sexes are weighted in the ratio of 2 males to 1 female. The reader may ask: Is this the weighting that we ought to have? The answer depends on the status of the interactions. If interactions are negligible, *any* weighting, provided that it is the same for every breed, furnishes unbiased estimates of the population differences between breed means. The 2:1 weighting gives the most precise estimates from the available data. If interactions are present, breed differences are not the same for males as for females, so that different weightings produce real differences in results. Usually, as emphasized on several occasions, we do not wish to examine main effects when interactions are present. If we do, a 2:1 weighting is appropriate, when interactions are present, only if it represents the proportions in which males and females appear in the target population of the study, as happens in this example. Equal weighting or some other proportion would be preferred if it were more typical of the population about which the investigator wishes to draw conclusions.

With unequal sub-class numbers the expressions for the expected values of the mean squares in terms of components of variance are complicated. Wilk and Kempthorne (3) have developed formulas for 2- and 3-factor arrangements: the sub-class numbers may be equal or proportional. With 2 factors, let the proportions in factor A be $u_1:u_2:\dots:u_a$ and those in B , $v_1:v_2:\dots:v_b$. The number of observations in the (i, j) sub-class will then be some multiple of $u_i v_j$, say $nu_i v_j$. Note the value of n . The mathematical model is as given in 16.2.1, where the α_i , β_j and

TABLE 16.4.1
DRESSING PERCENTAGES (LESS 70%) OF 93 SWINE CLASSIFIED BY BREED AND SEX
LIVE WEIGHTS 200-219 POUNDS

Number	Breed									
	1		2		3		4		5	
	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
1	13.3	18.2	10.9	14.3	13.6	12.9	11.6	13.8	10.3	12.8
2	12.6	11.3	3.3	15.3	13.1	14.4	13.2	14.4	10.3	8.4
3	11.5	14.2	10.5	11.8	4.1		12.6	4.9	10.1	10.6
4	15.4	15.9	11.6	11.0	10.8		15.2		6.9	13.9
5	12.7	12.9	15.4	10.9			14.7		13.2	10.0
6	15.7	15.1	14.4	10.5			12.4		11.0	
7	13.2		11.6	12.9					12.2	
8	15.0		14.4	12.5					13.3	
9	14.3		7.5	13.0					12.9	
10	16.5		10.8	7.6					9.9	
11	15.0		10.5	12.9						
12	13.7		14.5	12.4						
13			10.9	12.8						
14			13.0	10.9						
15			15.9	13.9						
16			12.8							
17			14.0							
18			11.1							
19			12.1							
20			14.7							
21			12.7							
22			13.1							
23			10.4							
24			11.9							
25			10.7							
26			14.4							
27			11.3							
28			13.0							
29			12.7							
30			12.6							
ΣX	168.9	87.6	362.7	182.7	41.6	27.3	79.7	33.1	110.1	55.7

Total. $N = 93$, $\Sigma X = 1,149.4$, $\Sigma X^2 = 14,785.62$

Breed Sums: 1, 256.5, 2, 545.4, 3, 68.9, 4, 112.8, 5, 165.8

Sex Sums: Male, 763.0, Female, 386.4

1 Correction $C = (\Sigma X)^2 / N = (1,149.4)^2 / 93 = 14,205.60$

2 Total $\Sigma X^2 - C = 14,785.62 - 14,205.60 = 580.02$

3 Sub-classes $\frac{(168.9)^2}{12} + \frac{(87.6)^2}{6} + \frac{(55.7)^2}{5} - C = 122.83$

4 Within sub-classes $580.02 - 122.83 = 457.19$

5 Sex $\frac{(763.0)^2}{62} + \frac{(386.4)^2}{31} - C = 0.52$

6 Breeds $\frac{(256.5)^2}{18} + \frac{(165.8)^2}{15} - C = 97.38$

7 Interaction $122.83 - (97.38 + 0.52) = 24.93$

	Degrees of Freedom	Sum of Squares	Mean Square
Sex	1	0.52	0.52
Breeds	4	97.38	24.34**
Breed-Sex Interactions	4	24.93	6.23
Within sub-classes	83	457.19	5.51

Breed Mean Percentages					
	1	2	3	4	5
	84.2	82.1	81.5	82.5	81.1
n_i .	18	45	6	9	15

I_{ij} may be either fixed or random. Also let:

$$U = \Sigma u, \quad V = \Sigma v, \quad U^* = \frac{\Sigma u^2}{(\Sigma u)^2}, \quad V^* = \frac{\Sigma v^2}{(\Sigma v)^2}$$

The expected values of the mean squares are:

$$E(A) = \sigma^2 + \frac{nUV(1 - U^*)}{a - 1} \left\{ \left(V^* - \frac{1}{b} \right) \sigma_{AB}^2 + \sigma_A^2 \right\}$$

$$E(B) = \sigma^2 + \frac{nUV(1 - V^*)}{b - 1} \left\{ \left(U^* - \frac{1}{a} \right) \sigma_{AB}^2 + \sigma_B^2 \right\}$$

$$E(AB) = \sigma^2 + \frac{nUV(1 - U^*)(1 - V^*)}{(a - 1)(b - 1)} \sigma_{AB}^2$$

These results hold when both factors are fixed. If A is random, delete the term in $1/a$ (inside the curly bracket) in $E(B)$. If B is random, delete the term in $1/b$ in $E(A)$. With fixed factors, the variance components are defined as follows:

$$\sigma_A^2 = \Sigma \alpha_i^2 / (a - 1) : \sigma_B^2 = \Sigma \beta_j^2 / (b - 1) : \sigma_{AB}^2 = \Sigma I_{ij}^2 / (a - 1)(b - 1)$$

For the example, if A denotes sex and B denotes breed:

$$a = 2, b = 5; u_1 = 2, u_2 = 1; v_1 = 6, v_2 = 15, v_3 = 2, v_4 = 3, v_5 = 5; n = 1$$

$$U = 3; V = 31; U^* = \frac{2^2 + 1^2}{3^2} = 0.556; V^* = \frac{6^2 + \dots + 5^2}{31^2} = 0.311$$

Regarding sex and breed as fixed parameters, we find

$$E(A) = \sigma^2 + 4.58\sigma_{AB}^2 + 41.3\sigma_A^2$$

$$E(B) = \sigma^2 + 0.90\sigma_{AB}^2 + 16.0\sigma_B^2$$

$$E(AB) = \sigma^2 + 7.11\sigma_{AB}^2$$

Note that $E(A)$ and $E(B)$ contain terms in the interaction variance, even though all effects are fixed. This happens because when the numbers are proportional, the main effects are *weighted* means. Although the I_j sum to zero over any row or column, their weighted means are not zero. As a further illustration, you may verify that if A were random in these data, we would have:

$$E(B) = \sigma^2 + 8.90\sigma_{AB}^2 + 16.0\sigma_B^2$$

Our second example (table 16.4.2) illustrates the use of analysis by proportional numbers as an approximation to the least squares analysis. In a sample survey of farm tenancy in an Iowa county (4), it was found that farmers had about the same proportions of Owned, Rented, and Mixed

TABLE 16.4.2
FARM ACRES IN CORN CLASSIFIED BY TENURE AND SOIL PRODUCTIVITY
AUDUBON COUNTY, IOWA

Soil Class		Owner		Renter		Mixed		Σn	ΣX
		Observed	Proportional	Observed	Proportional	Observed	Proportional		
I	n	36	36.75	67	62.92	49	52.33	152	7,323
	\bar{X}	32.7		55.2		50.6			
	ΣX		1,202		3,473		2,648		
II	n	31	33.85	60	57.95	49	48.20	140	6,584
	\bar{X}	36.0		53.4		47.1			
	ΣX		1,219		3,095		2,270		
III	n	58	54.40	87	93.13	80	77.47	225	9,102
	\bar{X}	30.1		46.8		40.1			
	ΣX		1,637		4,358		3,107		
Σn		125		214		178		517	
ΣX			4,058		10,926		8,025		23,009

Analysis of Variance Using Proportional Numbers

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Soils	2	6,635	3,318*
Tenures	2	27,367	13,684**
Interactions	4	883	221
Error (from original data)	508		830

Means

Owner	Renter	Mixed
32.5	51.1	45.1
I 48.2	II 47.0	III 40.5

farms in 3 soil fertility classes (section 9.13). Replacement of the actual sub-class numbers by numbers that are proportional should therefore give a good approximation to the least squares analysis. The proportional numbers are calculated from the row and column totals of the actual numbers. Thus, for Renters in Soil Class III, $93.13 = (225)(214)/517$. The sub-class means are multiplied by these fictitious numbers to produce the sub-class totals ΣX in table 16.4.2.

The variable being analyzed is the number of acres of corn per farm. There are large differences between tenure means, renters and mixed owner-renter farmers having more corn than owners. The amount of corn is also reduced on Soil Class III. There is no evidence of interactions. Since the proportional numbers agree so well with the actual numbers, an exact least squares analysis in these data is unnecessary. In general, analysis by proportional numbers should be an adequate approximation to the least-squares analysis if the ratios of the proportional to the actual cell numbers all lie between 0.75 and 1.3, although this question has not been thoroughly studied.

16.5—Disproportionate numbers. The 2×2 table. In section 16.7 the analysis of the $R \times C$ table when sub-class numbers are neither equal nor proportional will be presented. The 2×2 and the $R \times 2$ table, which are simpler to handle and occur frequently, are discussed in this and the next section. Table 16.5.1 gives an example (5). The data relate to the effects of two hormones on the comb weights of chicks.

TABLE 16.5.1
COMB WEIGHTS (MG.) OF LOTS OF CHICKS INJECTED WITH TWO SEX HORMONES

	Number	Untreated ΣX	\bar{X}	Number	Hormone A ΣX	\bar{X}
Untreated	3	240	80	12	1,440	120
Hormone B	12	1,200	100	6	672	112

The Within-classes mean square, computed from the individual observations, was $s^2 = 811$, with 29 *d.f.* To test the interaction, compute it from the sub-class means in the usual way for a 2×2 factorial:

$$80 + 112 - 100 - 120 = -28$$

Taking account of the sub-class numbers, the standard error of this estimate is

$$\sqrt{s^2 \left(\frac{1}{3} + \frac{1}{6} + \frac{1}{12} + \frac{1}{12} \right)} = \sqrt{(811) \frac{2}{3}} = \pm 23.25$$

The value of *t* is $-28/23.25 = -1.20$, with 29 *d.f.*, *P* about 0.25. We shall assume interaction unimportant and proceed to compute the main effects (table 16.5.2).

TABLE 16 5.2
CALCULATION OF MAIN EFFECTS OF HORMONES A AND B

	Untreated		Hormone A		$D_A =$	$W_A =$	$W_A D_A$
	n_1	\bar{X}_o	n_2	\bar{X}_A	$\bar{X}_A - \bar{X}_o$	$\frac{n_1 n_2}{n_1 + n_2}$	
Untreated	3	80	12	120	40	2.4	96
	12	100	6	112	12	4.0	48
Hormone B							
	W_B	D_B	W_B	D_B		6.4	144
	2.4	20	4.0	-8			

Main effect of A: $\Sigma W_A D_A / \Sigma W_A = 144 / 6.4 = 22.5$

$$S.E. = \sqrt{s^2 / \Sigma W_A} = \sqrt{811 / 6.4} = \pm 11.26 \text{ (29 d.f.)}$$

Main effect of B: $\Sigma W_B D_B / \Sigma W_B = 16 / 6.4 = 2.5$

$$S.E. = \sqrt{s^2 / \Sigma W_B} = \sqrt{811 / 6.4} = \pm 11.26 \text{ (29 d.f.)}$$

Consider Hormone A. The differences D_A between the means with and without A are recorded separately for the two levels of B. These are the figures 40 and 12. Since interaction is assumed absent, each figure is an estimate of the main effect of A. But the estimates differ in precision because of the unequal sub-class numbers. For an estimate derived from two sub-classes with numbers n_1 and n_2 the variance is

$$\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = \sigma^2 \frac{(n_1 + n_2)}{n_1 n_2}$$

Consequently, the estimate receives a relative weight $W = n_1 n_2 / (n_1 + n_2)$. These weights are computed and recorded. The main effect of A is the weighted mean of the two estimates, $\Sigma W D / \Sigma W$, with *s.e.* $\pm \sqrt{s^2 / \Sigma W}$. The main effect of B is computed similarly. The increase in comb weights due to Hormone A is 22.5 mg. ± 11.26 mg., almost significant at the 5% level, but Hormone B appears to have little effect.

Note: in this example the two values of W , 2.4 and 4.0, happen to be the same for A and B. This arises because two sub-classes are of size 12 and is not generally true. We have not described the analysis of variance because it is not needed.

16.6—Disproportionate numbers. The $R \times 2$ table. The data in table 16.6.1 illustrate some of the peculiarities of disproportionate sub-class numbers (6). In a preliminary analysis of variance, shown under the table, the Total sum of squares between sub-class means and the sums of squares for Sexes and Generations were computed by the usual elementary methods (taking account of the differences in sub-class numbers). The Interactions sum of squares was then found to be

$$119,141 - 114,287 - 5,756 = -902$$

The Sexes and Generations S.S. add to more than the total S.S. between sub-classes. This is because differences between the Generation means are inflated by the inequality in the Sex means, and vice versa.

TABLE 16.6.1
NUMBER, TOTAL GAIN, AND MEAN GAIN IN WEIGHT OF WISTAR RATS (GMS. MINUS 100)
IN FOUR SUCCESSIVE GENERATIONS. GAINS DURING SIX WEEKS FROM 28 DAYS OF AGE

Generation	Male			Female			$W_j = \frac{n_1 n_2}{n_{1j} + n_{2j}}$	$D_j = \bar{X}_{1j} - \bar{X}_{2j}$	
	n_{1j}	X_{1j}	\bar{X}_{1j}	n_{2j}	X_{2j}	\bar{X}_{2j}			$W_j D_j$
1	21	1,616	76.95	27	257	9.52	11.81	67.43	796.35
2	15	922	61.47	25	352	14.08	9.38	47.39	444.52
3	12	668	55.67	23	196	8.52	7.89	47.15	372.01
4	7	497	71.00	19	129	6.79	5.12	64.21	328.76
							34.20		1,941.64

Preliminary Analysis of Variance

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Sexes	1	114,287	
Generations	3	5,756	
Interactions	3	-902(!)	
Between sub-classes	7	119,141	
Within sub-classes	141		409

Calculation of Adjusted Generation Means

Generation	$n_{.j}$	$X_{.j}$	$\bar{X}_{.j}$	Estimate of	Adjusted Mean
1	48	1,873	39.02	$\mu + \alpha_1 - \delta/16$	42.57
2	40	1,274	31.85	$\mu + \alpha_2 - \delta/8$	38.95
3	35	864	24.69	$\mu + \alpha_3 - 11\delta/70$	33.61
4	26	626	24.08	$\mu + \alpha_4 - 3\delta/13$	37.18

In any $R \times 2$ table the correct Interactions S.S. is easily computed directly. Calculate the observed sex difference D and its weight W separately for each generation (table 16.6.1). The Interactions S.S. (3 d.f.) is given by

$$\Sigma WD^2 - (\Sigma WD)^2 / \Sigma W = (67.43)(796.35) + \dots + (64.21)(328.76) \\ - (1,941.64)^2 / (34.20) = 3,181$$

The F -test of Interactions is $F = 1,060/409 = 2.59$, close to the 5% level. It looks as if the sex difference was greater in generations 1 and 4 than in generations 2 and 3. There is, however, no *a priori* reason to anticipate that the sex difference would change from generation to generation. Perhaps the cell means were affected by some extraneous sources of varia-

tion that did not contribute to the variation within cells. For illustration, we proceed to estimate main effects on the assumption that interactions are negligible.

The estimate of the sex difference in mean gain is

$$D = \Sigma W_j D_j / \Sigma W_j = 1,941.64 / 34.20 = 56.77 \text{ gms.}$$

$$\text{Its S.E. is } \sqrt{s^2 / \Sigma W} = \sqrt{409 / 34.2} = 3.46 \text{ gms}$$

To estimate the Generation effects, note that under the additive model the population means for males and females in Generation j may be written as follows

$$\text{Males: } \mu + \alpha_j + \frac{1}{2}\delta; \quad \text{Females: } \mu + \alpha_j - \frac{1}{2}\delta$$

where δ represents the sex difference, Males minus Females. We start with the unadjusted mean for each generation and adjust it so as to remove the sex effect. Since generation 1 has 21 males and 27 females out of 48, its unadjusted mean is an unbiased estimate of

$$\mu + \alpha_1 + \frac{21}{48}\left(\frac{\delta}{2}\right) + \frac{27}{48}\left(-\frac{\delta}{2}\right) = \mu + \alpha_1 - \frac{\delta}{16}$$

Our estimate of δ is 56.77 and the unadjusted mean for generation 1 is 39.02. To remove the sex effect, we add $56.77/16 = 3.55$, giving 42.57. These adjustments are made at the foot of table 16.6.1.

For comparisons among these adjusted generation means, standard errors may be needed. The difference between the adjusted means of the j th and k th generation is of the form

$$\bar{X}_{.j} - \bar{X}_{.k} + gD,$$

where g is the numerical multiplier of D . The variance of this difference is

$$s^2 \left(\frac{1}{n_j} + \frac{1}{n_k} + \frac{g^2}{\Sigma W} \right)$$

With generations 1 and 2, $n_1 = 48$, $n_2 = 40$, while $g = (-1/16) - (-1/8) = 1/16$, and $\Sigma W = 34.2$. The term in g in the variance turns out to be negligible. The variance of the difference is therefore

$$(409) \left(\frac{1}{48} + \frac{1}{40} \right) = 18.73$$

The adjusted difference is 3.62 ± 4.33

If F -tests of the main effects of Sexes and Generations are wanted, start with the preliminary $S.S.$ for each factor in table 16.6.1. Subtract from it the difference:

Correct Interaction $S.S.$ minus Preliminary Interaction $S.S.$

$$= 3,181 - (-902) = 4,083$$

The resulting adjusted $S S$ are shown in table 16.6.2.

TABLE 16.6.2
ADJUSTED SUMS OF SQUARES OF MAIN EFFECTS

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Square
Sexes (adjusted)	1	$114,287 - 4,083 = 110,204$	110,204**
Generations (adjusted)	3	$5,756 - 4,083 = 1,673$	558
Interactions	3	3,181	1,060
Within sub-classes	141		409

The sex difference is large, but the generation differences fall short of the 5% level.

If interactions can be neglected, this analysis is applicable to tables in which data are missing entirely from a sub-class. Zeros are entered for the missing n_{ij} and $\bar{X}_{i\cdot}$. From the $d.f.$ for Interactions deduct 1 for each missing cell.

EXAMPLE 16.6.1—(i) Verify from table 16.6.2 that the adjusted $S S$ for Sexes, Generations, and Interactions do not add up to the Total $S S$ between sub-classes. (ii) Verify that the adjusted $S S$ for Sexes, 110,204, can be computed directly (apart from rounding errors), as $(\sum WD)^2 / \sum W$. This formula holds in all $R \times 2$ tables.

An additive analysis of variance can be obtained from the Preliminary $S S$ for Generations and the adjusted $S S$ for Sexes, as follows

	Degrees of Freedom	Sum of Squares
Generations (ignoring Sexes)	3	5,756
Sexes (adjusted for Generations)	1	110,204
Interactions	3	3,181
Total between sub-classes	7	119,141

This breakdown is satisfactory when we are interested only in testing Sexes. Alternatively, we can get an additive breakdown from Sexes (ignoring Generations) and Generations (adjusted).

EXAMPLE 16.6.2—Becker and Hall (10) determined the number of oocysts produced by rats of five strains during immunization with *Eimeria murrayi*. The unit of measurement is 10^6 oocysts.

Sex		Strain				
		Lambert	Lo	Hi	W E L	Wistar (A)
Male	n	8	14	20	8	9
	\bar{X}	36.1	94.9	194.4	64.1	175.7
Female	n	7	14	21	10	8
	\bar{X}	31.9	68.6	187.3	89.2	148.4

Verify the completed analysis of variance quoted from the original article:

Sex (adjusted)	1	2,594.6	2,594.6
Strains (adjusted)	4	417,565.6	104,391.4
Interaction	4	8,805.3	2,201.3
Within sub-classes*	109	332,962.9	3,054.7

* You cannot, of course, verify this line.

You will not be able to duplicate these numbers exactly because the means are reported to only 3 significant digits. Your results should approximate the first 3 figures in the mean squares, enough for testing.

16.7—The $R \times C$ table. Least squares analysis. This is a general method for analyzing 2-way classifications (7). It fits an additive model (i.e., one assuming no interactions) to the sub-class means:

$$\bar{X}_{ij} = \mu + \alpha_i + \beta_j + \bar{\epsilon}_{ij}, \quad i = 1, \dots, r, j = 1, \dots, c,$$

where the $\bar{\epsilon}_{ij}$ are assumed normally distributed with means zero and variances σ^2/n_{ij} , where n_{ij} is the sub-class number. This amounts to assuming that the variance within each subclass is σ^2 , since $\bar{\epsilon}_{ij}$ is the mean of n_{ij} such residuals.

As an intermediate step in the calculations, the method provides the most powerful test of the null hypothesis that interactions are zero. If this hypothesis is contradicted by the data, the calculations are usually stopped and the investigator proceeds to examine the two-way table in detail. If the assumption of negligible interactions is tenable, the remainder of the calculations give unbiased estimates of the row and column main effects α_i and β_j that have the smallest variances. Since data of this type are common and are tedious to handle on a desk machine, most computing centers are likely to have a standard program for the analysis.

The basic data used are the n_{ij} and the row ($X_{i\cdot}$) and column ($X_{\cdot j}$) totals of the observations. Table 16.7.1 shows the algebraic notation and the mouse typhoid data of table 16.2.1 used as illustration. (The p_{ij} are explained later.) Following Yates (7), we denote the row and column totals of the n_{ij} by $N_{i\cdot}$ and $N_{\cdot j}$.

The least squares method chooses estimates m, a_i, b_j of μ, α_i, β_j that minimize

$$\sum_i \sum_j n_{ij} (\bar{X}_{ij} - m - a_i - b_j)^2$$

The resulting normal equation for a_i is

$$N_{i\cdot}(m + a_i) + n_{i1}b_1 + n_{i2}b_2 + \dots + n_{ic}b_c = X_{i\cdot} \quad (16.7.1)$$

Thus, for Organism 9D, we have

$$98(m + a_1) + 34b_1 + 31b_2 + 33b_3 = 385$$

Note that the least squares method estimates α_i , the effect of the i th row, by making the observed total for the i th row equal the value which the model says it ought to have. Similarly, for the j th column,

$$N_{\cdot j}(m + b_j) + n_{1j}a_1 + n_{2j}a_2 + \dots + n_{rj}a_r = X_{\cdot j} \quad (16.7.2)$$

TABLE 16.7.1
ALGEBRAIC NOTATION AND DATA FOR FITTING THE ADDITIVE MODEL

	Columns			Totals	Data Totals	
	1	2	C			
n_{1j}	n_{11}	n_{12}	n_{1c}	$N_{1\cdot}$	$X_{1\cdot}$	
$p_{1j} = n_{1j}/N_{1\cdot}$	p_{11}	p_{12}	p_{1c}	1		
n_{2j}	n_{21}	n_{22}	n_{2c}	$N_{2\cdot}$	$X_{2\cdot}$	
$p_{2j} = n_{2j}/N_{2\cdot}$	p_{21}	p_{22}	p_{2c}	1		
n_{rj}	n_{r1}	n_{r2}	n_{rc}	N_r	X_r	
$p_{rj} = n_{rj}/N_r$	p_{r1}	p_{r2}	p_{rc}	1		
Data totals	$N_{\cdot 1}$ $X_{\cdot 1}$	$N_{\cdot 2}$ $X_{\cdot 2}$	$N_{\cdot c}$ $X_{\cdot c}$	N		X
Organism	Strain of Mice			$N_{i\cdot}$	$X_{i\cdot}$	\bar{X}_i
	RI	Z	Ba			
9D n_{1j}	34	31	33	98	385	3.929
p_{1j}	0.34694	0.31633	0.33673	1		
11C n_{2j}	66	78	113	257	1,442	5.611
p_{2j}	0.25681	0.30350	0.43969	1		
DSCI n_{3j}	107	133	188	428	2,523	5.895
p_{3j}	0.25000	0.31075	0.43925	1		
$N_{\cdot j}$	207	242	334	783		
$X_{\cdot j}$	1,271	1,692	1,387		4,350	5.556
b_j	2.1251	2.8986	0			

From (16.7.1) we see that if we know the b 's, we can find $(m + a_i)$, while if we know the a 's, (16.7.2) gives $(m + b_j)$. The next step is to eliminate either the a 's or the b 's. Time is usually saved by eliminating the more numerous set of constants, though an investigator interested only in the rows may prefer to eliminate the columns. In this example, with $r = c = 3$, it makes no difference. We shall eliminate the a 's (rows).

When the a 's are eliminated, m also disappears. In finding the equations for the b 's, it helps to divide each n_{ij} by its row total $N_{i\cdot}$, forming the p_{ij} . The equations for the b 's are derived by a rule that is easily remembered. The first equation is

$$(N_{\cdot 1} - n_{11}p_{11} - \dots - n_{r1}p_{r1})b_1 - (n_{11}p_{12} + \dots + n_{r1}p_{r2})b_2 - \dots \\ - (n_{11}p_{1c} + \dots + n_{r1}p_{rc})b_c = X_{\cdot 1} - p_{11}X_{1\cdot} - \dots - p_{r1}X_r.$$

For the mice, the first equation is

$$[207 - (34)(0.34694) - \dots - (107)(0.25000)]b_1 \\ - [(34)(0.31633) + \dots + (107)(0.31075)]b_2 \\ - [(34)(0.33673) + \dots + (107)(0.43925)]b_3 \\ = 1,271 - (0.34694)(385) - \dots - (0.25000)(2,523)$$

In the j th equation the term in b_j is $N_{\cdot j}$ minus the sum of products of the

n 's and p 's in that column. The term in b_k is *minus* the sum of products of the n_{ij} and the p_{ik} . The three equations are:

$$\begin{aligned} 151.505b_1 - 64.036b_2 - 87.468b_3 &= 136.35 \\ -64.036b_1 + 167.191b_2 - 103.155b_3 &= 348.54 \\ -87.468b_1 - 103.155b_2 + 190.624b_3 &= -484.92 \end{aligned} \quad (16.7.2a)$$

The sum of the numbers in each of the four columns above adds to zero, apart from rounding errors. This is a useful check.

In previous analyses of 2-way tables in this book, we have usually assumed $\Sigma b_i = 0$. In solving these equations it is easier to assume $b_3 = 0$. (This gives exactly the same results for any comparison among the b 's.) Drop b_3 from the first two equations and drop the third equation, solving the equations

$$\begin{aligned} 151.505b_1 - 64.036b_2 &= 136.35 \\ -64.036b_1 + 167.191b_2 &= 348.54 \end{aligned}$$

The inverse of the 2×2 matrix (section 13.4) is

$$\begin{pmatrix} 0.0078753 & 0.0030163 \\ 0.0030163 & 0.0071365 \end{pmatrix} \quad (16.7.3)$$

giving

$$b_1 = 2.1251 : b_2 = 2.8986 : b_3 = 0 \quad (16.7.4)$$

The sum of squares for columns, adjusted for rows, is given by the sum of products of the b 's with the right sides of equations (16.7.2a).

$$\text{Column S.S. (adjusted)} = (2.1251)(136.35) + (2.8986)(348.54) = 1,300$$

The analysis of variance can now be completed and the Interactions S.S. tested. Compute the S.S. Between sub-classes and the unadjusted S.S. for Rows and Columns, these being, respectively,

$$\Sigma \Sigma X_{ij}^2 / n_{i.} - C; \Sigma X_{i.}^2 / N_{i.} - C; \Sigma X_{.j}^2 / N_{.j} - C;$$

where $C = X_{...}^2 / N_{...}$. The results are shown in the top half of table 16.7.2.

In the completed analysis of variance, the S.S. Between sub-classes, 1,786, can be partitioned either into

Rows S.S. (unadjusted) + Columns S.S. (adjusted) + Interactions S.S.
or into

Rows S.S. (adjusted) + Columns S.S. (unadjusted) + Interactions S.S.

Since we now know that Rows S.S. (unadjusted) = 309 and Columns S.S. (adjusted) = 1,300, the first of these relations gives the Interactions S.S. as

$$1,786 - 309 - 1,300 = 177$$

The *d.f.* are $(r - 1)(c - 1) = 4$ in this example. The second relation provides the Rows *S.S.* (adjusted). The completed analysis of variance appears in the lower half of table 16.7.2.

TABLE 16.7.2
ANALYSIS OF VARIANCE OF THE MICE DATA

Source of Variation	Preliminary (Unadjusted)		
	Degrees of Freedom	Sum of Squares	Mean Square
Between sub-classes	8	1,786	
Rows (Organisms), unadjusted	2	309	
Columns (Strains), unadjusted	2	1,227	
Completed			
Rows (Organisms), unadjusted	2	309	650.0
Columns (Strains), adjusted	2	1,300	
Rows (Organisms), adjusted	2	382	191.0
Columns (Strains), unadjusted	2	1,227	
Interactions	4	177	44.2
Within sub-classes	774		5.015

As in the approximate analyses, interactions are shown to be present so that ordinarily the analysis would not be carried further; the data would be interpreted as in section 16.2. But to illustrate the computations we proceed as though there were no interaction. The mean squares for *F*-tests of the main effects of rows and columns are the *adjusted* mean squares in table 16.7.2.

The standard error of any comparison $\Sigma L_j b_j$ among the column main effects is

$$s\sqrt{(\Sigma L_j^2 c_{jj} + 2\Sigma L_j L_k c_{jk})}$$

where $s = \sqrt{5.015} = 2.24$ and the c_{jk} are the inverse multipliers in (16.7.3). Since b_3 was arbitrarily made 0, all c_{j3} are 0. As examples,

$$S.E.(b_1 - b_2) = 2.24\sqrt{.00788 + .00714 - 2(.00302)} = \pm 0.212$$

$$S.E.(b_1 - b_3) = 2.24\sqrt{.00788} = \pm 0.199$$

The row main effects can be obtained from (16.7.1), rewritten as

$$m + a_i = \bar{X}_{i..} - p_{i1}b_1 - \dots - p_{ic}b_c \quad (16.7.5)$$

In table 16.7.1, the $\bar{X}_{i..}$ are in the right-hand column and the b_j are at the foot of each column. Relation (16.7.5) gives

$$m + a_1 = 3.929 - (0.34694)(2.1251) - (0.31633)(2.8986) = 2.275$$

Similarly, we find

$$m + a_2 = 4.186 \quad : \quad m + a_3 = 4.463$$

From (16.7.5) any comparison $\Sigma L_i(m + a_i)$ among the row means is of the form

$$\Sigma L_i \bar{X}_{i..} - \Sigma u_j b_j$$

To find the variance of this comparison, multiply s^2 by

$$\Sigma \frac{L_i^2}{N_{i.}} + \Sigma u_j^2 c_{jj} + 2 \Sigma u_j u_k c_{jk}$$

For example, the difference $a_2 - a_1 = 1.911$, is

$$\bar{X}_{2..} - \bar{X}_{1..} + 0.0901b_1 + 0.0128b_2$$

The multiplier of s^2 is, therefore,

$$\begin{aligned} & \frac{1}{98} + \frac{1}{257} \\ & + (0.0901)^2(0.00788) + (0.0128)^2(0.00714) \\ & + 2(0.0901)(0.0128)(0.00302) = 0.01417 \end{aligned}$$

The *s.e.* is $\pm \sqrt{(5.015)(0.01417)} = \pm 0.266$.

In the original data the overall mean is $\bar{X}... = 4350/783 = 5.556$ (table 16.7.1). Our three estimated row means are all less than 5.556. This is a consequence of the choice of $b_3 = 0$ to simplify the arithmetic. Although this choice has no effect on any comparison among the row means $m + a_i$ or the column means $m + b_j$, it is sometimes desirable to adjust the $m + a_i$ and the $m + b_j$ so that m becomes $\bar{X}...$. To do this, calculate the weighted mean of the $m + a_i$ with weights $N_{i.}$; that is,

$$[(98)(2.275) + (257)(4.186) + (428)(4.463)]/783 = 4.098$$

Since $\bar{X}... = 5.556$, we add +1.458 to each $m + a_i$, giving 3.733, 5.644, and 5.921 for the row means. To make the column means average in the same way to the general mean, compute these means as $\bar{X}... + b_j - 1.458$, giving the values 6.223, 6.997, 4.098.

In a 3-way classification the exact methods naturally become more complicated. There are now three two-factor interactions and 1 three-factor interaction. An example worked in detail is given by Stevens (8).

The exact analysis of variance can still be computed by elementary methods if the sub-class numbers are proportional, that is, if

$$n_{ijk} = (N_{i..})(N_{.j.})(N_{..k})/N...^2$$

Federer and Zelen (9) present exact methods for computing the sum of squares due to any main effect or interaction, assuming all other effects present. They also describe simpler methods that provide close upper

bounds to these sums of squares. Their methods are valid for any number of classes.

EXAMPLE 16.7.1—In the farm tenancy example in section 16.4 there was no evidence of interaction. The following are the least squares estimates of the main effect means for tenure and soils.

Owner:	32.507	Renter:	51.072	Mixed:	45.031
I :	48.157	II :	46.999	III :	40.480

Your results may differ a little, depending on the number of decimals carried. The results above were adjusted so that $\sum N_{i.}a_i = \sum N_{.j}b_j = 0$. Note the excellent agreement given by the means shown under table 16.4.2 for the method of proportional numbers

EXAMPLE 16.7.2—In the mice data verify the following estimates and standard errors as given by the use of equal weights within rows (section 16.3) and the least squares analysis (section 16.7).

	Equal Within Rows	Least Squares
11C - 9D	1.919 ± 0.265	1.911 ± 0.266
Z - R1	0.755 ± 0.196	0.774 ± 0.212

16.8—The analysis of proportions in 2-way tables. In chapter 9 we discussed methods of analysis for a binomial proportion. Sections 9.8–9.11 dealt with a set of C proportions arranged in a one-way classification. Two-way tables in which the entry in every cell is a sample proportion are also common. Examples are sample survey results giving the percentage of voters who stated their intention to vote Democratic, classified by the age and income level of the voter, or a study of the proportion of patients with blood group O in a large hospital, classified by sex and type of illness.

The data consist of rc independent values of a binomial proportion $p_{ij} = g_{ij}/n_{ij}$, arranged in r rows and c columns. The data resemble those in the preceding section, but instead of having a sample of continuous measurements X_{ijk} ($k = 1, 2, \dots, n_{ij}$) in the i, j cell, we have a binomial proportion p_{ij} . The questions of interest are usually the same in the binomial and the continuous cases. We want to examine whether row and column effects are additive, and if so, to estimate them and make comparisons among rows and among columns. If interactions are present, the nature of the interactions is studied.

From the viewpoint of theory, the analysis of proportions presents more difficulties than that of normally distributed continuous variables. Few exact results are available. The approximate methods used in practice mostly depend on one of the following approaches.

1. Regard p_{ij} as a normally distributed variable with variance $p_{ij}q_{ij}/n_{ij}$, using the weighted methods of analysis in preceding sections, with weights $w_{ij} = n_{ij}/p_{ij}q_{ij}$ and p_{ij} replacing \bar{X}_{ij} .

2. Transform the p_{ij} to equivalent angles y_{ij} (section 11.16), and treat the y_{ij} as normally distributed. Since the variance of y_{ij} for any p_{ij} is approximately $821/n_{ij}$, this method has the advantage that if the n_{ij} are constant, the analysis of variance of the y_{ij} is unweighted. As we have seen, this transformation is frequently used in randomized blocks experiments in which the measurement is a proportion.

3. Transform p_{ij} to its logit $Y_{ij} = \log_e (p_{ij}/q_{ij})$. The estimated variance of Y_{ij} is approximately $1/(n_{ij}p_{ij}q_{ij})$, so that in a logit analysis, Y_{ij} is given a weight $n_{ij}p_{ij}q_{ij}$.

The assumptions involved in these approaches probably introduce little error in the conclusions if the observed numbers of successes and failures, $n_{ij}p_{ij}$ and $n_{ij}q_{ij}$, exceed 20 in every cell. Various small-sample adjustments have been prepared to extend the validity of the methods.

When all p_{ij} lie between 25% and 75%, the results given by the three approaches seldom differ materially. If, however, the p_{ij} cover a wide range from close to zero up to 50% or beyond, there are reasons for expecting that row and column effects are more likely to be additive on a logit scale than on the original p scale. To repeat an example cited in section 9.14, suppose that the data are the proportions of cases in which the driver of the car suffered injury in automobile accidents classified by severity of impact (rows) and by whether the driver wore a seat belt or not (columns). Under very mild impacts p is likely to be close to zero for both wearers and non-wearers, with little if any difference between the two columns. At the other end, under extreme impacts, p will be near 100% whether a seat belt was worn or not, with again a small column effect. The beneficial effect of the belts, if any, will be revealed by the accidents that show intermediate proportions of injuries. The situation is familiar in biological assay in which the toxic or protective effects of different agents are being compared. It is well known that two agents cannot be effectively compared at concentrations for which p is close to zero or 100%; instead, the investigator aims at concentrations yielding p around 50%.

Thus, in the scale of p , row and column effects cannot be strictly additive over the whole range. The logit transformation pulls out the scale near 0 and 100%, so that the scale extends from $-\infty$ to $+\infty$. In the logit analysis row and column effects may be additive, whereas in the p scale for the same data we might have interactions that are entirely a consequence of the scale. The angular transformation occupies an intermediate position. As with logits, the scale is stretched at the ends, but the total range remains finite, from 0 to 90°.

To summarize, with an analysis in the original scale it is easier to think about the meaning and practical importance of effects in this scale. The advantage of angles is the simplicity of the computations if the n_{ij} are equal or nearly so. Logits may permit an additive model to be used in tables showing large effects. In succeeding sections some examples will be given to illustrate the computations for analyses in the original and the logit scales.

The preceding analyses utilize *observed weights*, the weight $W = n/pq$ attached to the proportion p in a cell being computed from the observed value of p . Instead, when fitting the additive model we could use *expected weights* $\bar{W} = n/\hat{p}\hat{q}$, where \hat{p} is the estimate given by the additive model. This approach involves a process of successive approximations. We guess first approximations to the weights and fit the model, obtaining second approximations to the \hat{p} . From these, the weights are recomputed and the model fitted again, giving third approximations to the \hat{p} and the \bar{W} , and so on until no appreciable change occurs in the results.

This series of calculations may be shown to give successive approximations to maximum likelihood estimates of the \hat{p} (11). When np and nq are large in the cells, analyses by observed and expected weights agree closely. In small samples it is not yet clear that either method has a consistent advantage. Observed weights require less computation.

A word of caution: we are assuming that in any cell there is a single *binomial proportion*. Sometimes the data in a cell come from several binomials with different p 's. In a study of absenteeism among clerical workers, classified by age and sex, the basic measurement might be the proportion of working days in a year on which the employee was absent. But the data in a cell, e.g., men aged 20–25, might come from 18 different men who fall into this cell. In this event the basic variable is p_{ijk} , the proportion of days absent for the k th man in the i, j cell. Usually it is adequate to regard p_{ijk} as a continuous variate, performing the analysis by the methods in preceding sections.

16.9—Analysis in the p scale: a 2×2 table. In this and the next section, two examples are given to illustrate situations in which direct analysis of the proportions is satisfactory. Table 16.9.1 shows data cited by Bartlett (12) from an experiment in which root cuttings of plum trees were planted as a means of propagation. The factors are length of cutting (rows) and whether planting was done at once or in spring (columns).

TABLE 16.9.1
PERCENTAGES OF SURVIVING PLUM ROOT-STOCKS FROM 240 CUTTINGS

Length of Cutting	Time of Planting	
	At Once	Spring
Long	$p_{11} = 156/240 = 65.0\%$	$p_{12} = 84/240 = 35.0\%$
	$v_{11} = (65.0)(35.0)/240 = 9.48$	$v_{12} = (35.0)(65.0)/240 = 9.48$
Short	$p_{21} = 107/240 = 44.6\%$	$p_{22} = 31/240 = 12.9\%$
	$v_{21} = (44.6)(55.4)/240 = 10.30$	$v_{22} = (12.9)(87.1)/240 = 4.68$

In the (1, 1) cell, 156 plants survived out of 240, giving $p_{11} = 65.0\%$. The estimated variances v for each p are also shown.

The analysis resembles that of section 16.5, the p_{ij} replacing the \bar{X}_{ij} . To test for interaction we compute

$$p_{11} + p_{22} - p_{12} - p_{21} = 65.0 + 12.9 - 35.0 - 44.6 = -1.7\%$$

Its standard error is

$$\sqrt{(v_{11} + v_{22} + v_{12} + v_{21})} = \sqrt{33.94} = \pm 5.83$$

Since there is no indication of interaction, the calculation of row and column effects proceeds as in table 16.9.2. For the column difference in row 1, the variance is $(v_{11} + v_{12}) = 18.96$. The overall column difference is a weighted mean of the differences in the two rows, weighted inversely as the estimated variances. Both main effects are large relative to their standard errors. Clearly, long cuttings planted at once have the best survival rate.

TABLE 16.9.2
CALCULATION OF ROW AND COLUMN EFFECTS

	At Once	Spring	<i>D</i>	<i>V</i>	<i>W</i>
Long	$p_{11} = 65.0$ $v_{11} = 9.48$	$p_{12} = 35.0$ $v_{12} = 9.48$	30.0	18.96	0.0527
Short	$p_{21} = 44.6$ $v_{21} = 10.30$	$p_{22} = 12.9$ $v_{22} = 4.68$	31.7	14.98	0.0668
	$D = 20.4$ $V = 19.78$ $W = 0.0506$	$D = 22.1$ $V = 14.16$ $W = 0.0706$			

Main Effects:

At Once - Spring: $\Sigma WD/\Sigma W = 31.0\%$: $S.E. = 1/\sqrt{(\Sigma W)} = \pm 2.89$

Long - Short $\Sigma WD/\Sigma W = 21.4\%$: $S.E. = 1/\sqrt{(\Sigma W)} = \pm 2.87$

In Bartlett's original paper (12), these data were used to demonstrate how to test for interaction in the logit scale. (He regarded the data as a $2 \times 2 \times 2$ contingency table and was testing the three-factor interaction among the factors alive-dead, long-short, at once-spring.) However, the data show no sign of interaction either in the p or the logit scale.

16.10—Analysis in the p scale: a 3×2 table. In the second example, inspection of the individual proportions indicates interactions that are due to the nature of the factors and would not be removed by a logit transformation. The data are the proportions of children with emotional problems in a study of family medical care (13), classified by the number of children in the family and by whether both, one, or no parents were recorded as having emotional problems, as shown in table 16.10.1.

In families having one or no parents with emotional problems the four values of p are close to 0.3, any differences being easily accountable by sampling errors. Thus there is no sign of an effect of number of children or of the parents' status when neither or only one parent has emotional problems. When both parents have problems there is a marked increase in p in the smaller families to 0.579 and a modest increase in the larger

TABLE 16.10.1
PROPORTION OF CHILDREN WITH EMOTIONAL PROBLEMS

Parents With Problems	Number of Children in Family	
	1-2	3-4
Both	$p = 33/57 = 0.579$	$p = 15/38 = 0.395$
One	$p = 18/54 = 0.333$	$p = 17/55 = 0.309$
None	$p = 10/37 = 0.270$	$p = 9/32 = 0.281$

families to 0.395. Thus, inspection suggests that the proportion of children with emotional problems is increased when both parents have problems, and that this increase is reduced in the larger families.

Consequently, the statistical analysis would probably involve little more than tests of the differences $(0.579 - 0.333)$, $(0.395 - 0.309)$, and $(0.579 - 0.395)$, which require no new methods. The first difference is significant at the 5% level but the other two are not, so that any conclusions must be tentative. In data of this type nothing seems to be gained by transformation to logits. Reference (13) presents additional data bearing on the scientific issue.

16.11—Analysis of logits in an $R \times C$ table. When the fitting of an additive model in the logit scale is appropriate, the following procedure should be an adequate approximation:

1. If p is a binomial proportion obtained from g successes out of n trials in a typical cell of the 2-way table, calculate the logit as

$$Y = \ln\{(g + 1/2)/(n - g + 1/2)\}$$

in each cell, where \ln denotes the log to base e .

2. Assign to the logit a weight $W = (g + 1/2)(n - g + 1/2)/(n + 1)$. In large samples, with all values of g and $(n - g)$ exceeding 30, Y will be essentially $\ln(p/q)$ and the weight npq , which may be used if preferred. The values suggested here for Y and W in small samples are based on research by Gart and Zweifel (14). See example 16.12.3.

3. Then follow the method of fitting described for continuous data in section 16.7, with Y_{ij} in place of \bar{X}_{ij} , and with W_{ij} in place of n_{ij} as weights. Section 16.7 should be studied carefully.

4. The analysis of variance of Y is like table 16.7.2, but has no "Within sub-classes" line. If the additive model fits, the Interactions sum of squares is distributed approximately as χ^2 with $(r - 1)(c - 1)$ d.f. A significant value of χ^2 is a sign that the model does not fit. This test should be supplemented by inspection of the deviations $Y_{ij} - \hat{Y}_{ij}$ to note any systematic pattern that suggests that the model distorts the data.

5. If the model fits and the inverse multipliers c_{ij} have been computed for the columns, the s.e. of any linear function of the column main effects is

$$\sqrt{(\sum L_j^2 c_{jj} + 2 \sum L_j L_k c_{jk})}$$

In the numerical example which follows, the proportions p are all small, the largest being 0.056. In this event, the logit of p is practically the same as $\ln(p)$. In effect, we are fitting an additive model to the logarithms of the p 's, i.e., a multiplicative model to the p 's themselves. Further, with large samples the observed weight $W = npq$ becomes $W = np = g$ when p is small, each logit being weighted by the numerator of p .

16.12—Numerical example. The data come from a large study of the relationship between smoking and death rates (15). About 248,000 male policyholders of U.S. Government Life Insurance answered questions by mail about their smoking habits. The data examined here are for men who reported themselves as non-smokers and for men who reported that they smoked cigarettes only. The cigarette smokers are classified by number smoked per day, 1–9, 10–20, 21–39, and over 39. For each smoking class, the person-years of exposure were accumulated by 10-year age classes, using actual ages. That is, a man aged 52 on entry into the study would contribute 3 years in the 45–54 age class and additional years in the 55–64 age class. Most men were in the study for $8\frac{1}{2}$ years.

In table 16.12.1, part (A) shows for each cell the number of deaths. Part (B) gives the annual probability of death ($\times 10^3$) within each cell, calculated from the number of deaths and the number of person-years of exposure. Since the age distributions of different smoking classes were not identical within a 10-year age class, the probabilities were computed, by standard actuarial methods, so as to remove any effect of these differences in age-distributions.

TABLE 16.12.1
NUMBERS OF DEATHS AND ANNUAL PROBABILITIES OF DEATH ($\times 10^3$)

Age (Years)	Reported Number of Cigarettes Smoked Per Day				
	None	1–9	10–20	21–39	Over 39
<i>(A) number of deaths</i>					
35–44	47	7	90	83	10
45–54	38	11	67	80	14
55–64	2,617	389	2,117	1,656	406
65–74	3,728	586	2,458	1,416	258
<i>(B) annual probabilities of death ($\times 10^3$)</i>					
35–44	1.27	1.63	1.99	2.66	3.26
45–54	2.64	6.23	6.64	8.91	11.60
55–64	10.56	14.35	18.50	20.87	27.40
65–74	24.11	35.76	42.26	49.40	55.91

In every age group the probability of death rises sharply with each additional amount smoked. As expected, the probability also increases consistently with age within every smoking class. It is of interest to examine whether the rate of increase in probability of death for each additional

amount smoked is constant at the different ages or changes from age to age. If the rate of increase is constant, this implies a simple multiplicative model for row and column effects: apart from sampling errors, the probability p_{ij} for men in the i th age class and j th smoking class is of the form

$$p_{ij} = \alpha_i \beta_j$$

In natural logs this gives the additive model

$$\ln(p_{ij}) = \ln \alpha_i + \ln \beta_j$$

Before attempting to fit this model it may be well to compute for each age group the ratio of the smoker to the non-smoker probabilities of death (table 16.12.2) to see if the data seem to follow the model.

TABLE 16.12.2
RATIOS OF SMOKER TO NON-SMOKER PROBABILITIES OF DEATH

Age	λ_i	Reported Number Smoked Per Day			
		1-9	10-20	21-39	Over 39
35-44	11	1.28	1.57	2.09	2.57
45-54		2.36	2.51	3.37	4.39
55-64		1.36	1.75	1.98	2.59
65-74		1.48	1.75	2.05	2.32

The ratios agree fairly well for age groups 35-44, 55-64, and 65-74, but run substantially higher in age group 45-54. This comparison is an imprecise one, however, since the probabilities that provide the denominators for the ages 35-44 and 45-54 are based on only 47 and 38 deaths, respectively. A stabler comparison is given by finding in each row the simple average of the five probabilities and using this as denominator for the row. This comparison (example 16.12.1) indicates that the non-smoker probability of death may have been unusually low in the age group 45-54.

Omitting the multiplier 10^3 , the p values in table 16.12.1 range from 0.00127 to 0.05591. The assumption that these p 's are binomial is not strictly correct. Within an individual cell the probability of dying presumably varies somewhat from man to man. This variation makes the variance of p for the cell less than the binomial variance (see example 16.12.2), but with small p 's the difference is likely to be negligible. Further, as already mentioned, the p 's were adjusted in order to remove any difference in age distribution within a 10-year class. Assuming the p 's binomial, each $\ln p$ is weighted by the observed number of deaths in the cell, as pointed out at the end of the preceding section.

The model is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where the ε_{ij} are independent with means zero and variances $1/W_{ij}$. The fitted model is

$$\hat{Y}_{ij} = m + a_i + b_j,$$

the parameters being chosen so as to minimize $\sum W(Y - \hat{Y})^2$.

TABLE 16.12.3
ARRANGEMENT OF DATA FOR FITTING AN ADDITIVE MODEL

Age	Reported Number of Cigarettes Per Day						
		None	1-9	10-20	21-39	Over 39	
35-44	W_{1j} *	47	7	90	83	10	237 = $W_{1.}$
	Y_{1j} †	0.239	0.489	0.688	0.978	1.182	169.570 = $Y_{1.}$
45-54	W_{2j}	38	11	67	80	14	210
	Y_{2j}	0.971	1.829	1.893	2.187	2.451	393.122
55-64	W_{3j}	2.617	389	2,117	1,656	406	7,185
	Y_{3j}	2.357	2.664	2.918	3.038	3.310	19,756.759
65-74	W_{4j}	3,728	586	2,458	1,416	258	8,446
	Y_{4j}	3.183	3.577	3.744	3.900	4.024	29,725.690
	$W_{.j}$	6,430	993	4,732	3,235	688	16,078 = $W_{..}$
	$\bar{Y}_{.j}$	2.812	3.178	3.290	3.341	3.529	50,045.141 = Y 3.1126 = \bar{Y}

* W_{ij} = cell weight = number of deaths.

† $Y_{ij} = \ln(p_{ij})$.

Table 16.12.3 shows the weights W_{ij} = number of deaths and the $Y_{ij} = \ln(p_{ij})$. The first step is to find the row and column totals of the weights, and the weighted row and column totals of the Y_{ij} , namely

$$W_{.j} = \sum_i W_{ij} : W_{.j} = \sum_i W_{ij} : Y_{.j} = \sum_i W_{ij} Y_{ij} : Y_{.j} = \sum_i W_{ij} Y_{ij} :$$

$$W_{..} = \sum_i W_{i.} : Y_{..} = \sum_i Y_{i.}$$

If we make the usual restrictions,

$$\sum_i W_{i.} a_i = \sum_j W_{.j} b_j = 0,$$

then m is the overall mean $Y / W = \bar{Y} = 3.1126$. Analogous to (16.7.1) and (16.7.2), the normal equations for a_i and b_j are

$$W_{.1}(m + a_1) + W_{.2}b_2 + \dots + W_{.c}b_c = Y_{.1} \quad (16.12.1)$$

$$W_{.j}(m + b_j) + W_{.1}a_1 + W_{.2}a_2 + \dots + W_{.r}b_r = Y_{.j} \quad (16.12.2)$$

Since we are not interested in attaching standard errors to the a_i or b_j , these equations will be solved directly by successive approximations. As first approximations to the quantities $(m + b_j)$ we use the observed

column means $\bar{Y}_{.j} = Y_{.j}/W_{.j}$, shown in table 16.12.3. Rewriting equation (16.12.1) in the form

$$W_{i.}(m + a_i) = Y_{i.} + W_{i.}\bar{Y}_{..} - W_{i1}(m + b_1) - \dots - W_{ic}(m + b_c)$$

we obtain second approximations to the $(m + a_i)$. For row 1,

$$237(m + a_1) = 169.570 + (237)(3.1126) - (47)(2.812) - \dots - (10)(3.529)$$

$$(m + a_1) = 144.153/237 = 0.608$$

These are then inserted in (16.12.2) in the form

$$W_{.j}(m + b_j) = Y_{.j} + W_{.j}\bar{Y}_{..} - W_{1j}(m + a_1) - \dots - W_{rj}(m + a_r)$$

and so on. The estimates settle down quickly. After three rounds the following estimates were obtained:

Ages	35-44	45-54	55-64	65-74	
$m + a_i$	0.5748	1.7130	2.7193	3.5538	
No. per Day $m + b_j$	None 2.7433	1-9 3.1053	10-20 3.3052	21-39 3.4492	Over 39 3.6612

As a check, at each stage the quantities $\Sigma W_{i.}(m + a_i)$ and $\Sigma W_{.j}(m + b_j)$ should agree with the grand total $Y_{..}$ to within rounding errors.

The expected value in each cell is conveniently computed as

$$\hat{Y}_{ij} = (m + a_i) + (m + b_j) - \bar{Y}_{..}$$

Table 16.12.4 shows the observed and expected values and the deviations. The value of $\chi^2 = \Sigma W_{ij}(Y_{ij} - \hat{Y}_{ij})^2$ is 13.2 with 12 *d.f.*, giving no indication of a lack of fit. The largest deviation is the deficit -0.373 for non-smokers aged 45-54: this deviation also makes the largest contribution to χ^2 . The pattern of + and - signs in the deviations has no striking features.

By finding the antilogs of the quantities $(b_j - b_1)$, the ratios of the smoker to the non-smoker annual probabilities of death as given by this model are obtained. These ratios were 1.44, 1.75, 2.03, and 2.50, respectively, for smokers of 1-9, 10-20, 21-39, and over 39 cigarettes per day.

An example of the analysis of a proportion in a 2^4 factorial classification with only main effects important is given by Yates (16) using the logit scale and observed weights. Dyke and Patterson (17) give the maximum likelihood analysis of the same data. These authors define the logit as $\frac{1}{2}\ln(p/q)$.

Data containing a proportion in an $R \times C$ table may be regarded as an $R \times C \times 2$ contingency table, or as a particular case of an $R \times C \times T$ contingency table. The definition and testing of three-factor interactions

TABLE 16 12 4
OBSERVED AND EXPECTED NUMBERS OF $\ln p$

Age	Reported Number of Cigarettes Per Day					
	None	1-9	10-20	21-39	Over 39	
35-44	Y_{ij}	0 239	0 489	0 688	0 978	1 182
	\bar{Y}_{ij}	0 206	0 568	0 767	0 911	1 123
	D_{ij}	+0 033	-0 079	-0 079	+0 067	+0 059
45-54		0 971	1 829	1 893	2 187	2 451
		1 344	1 706	1 906	2 050	2 262
		-0 373	+0 123	-0 013	+0 137	+0 189
55-64		2 357	2 664	2 918	3 038	3 310
		2 350	2 712	2 912	3 056	3 268
		+0 007	-0 048	+0 006	-0 018	+0 042
65-74		3 183	3 577	3 744	3 900	4 024
		3 184	3 546	3 746	3 890	4 102
		-0 001	+0 031	-0 002	+0 010	-0 078

in such tables has attracted much attention in recent years. Goodman (18) gives a review and some simple computing methods.

EXAMPLE 16 12 1—In each row of table 16 12 1 find the unweighted average of the probabilities and divide the individual probabilities by this number. Show that the results are as follows.

Age	None	1-9	10-20	21-39	Over 39
35-44	59	75	92	1 23	1 51
45-54	37	86	92	1 24	1 61
55-64	58	78	1 01	1 14	1 49
65-74	58	86	1 02	1 19	1 35

The two numbers that seem most out of line are the low value 0 37 for (None 45-54) and the low value 1 35 for (over 39, 65-74).

EXAMPLE 16 12 2—Suppose that there are three groups of n men with probabilities of dying 0 01, 0 02, and 0 03. The variance of the total number who die is

$$n[(0 01)(0 99) + (0 02)(0 98) + (0 03)(0 97)] = 0 0586n$$

Hence the variance of the proportion of those dying out of $3n$ is $0 0586n/9n = 0 006511/n$. For the combined sample the probability of dying is 0 02. If we wrongly regard the combined sample as a single binomial of size $3n$ with $p = 0 02$, we would compute the variance of the proportion dying as $(0 02)(0 98)/3n = 0 006533/n$. The actual variance is just a trifle smaller than the binomial variance.

If there are k groups of men with probabilities p_1, p_2, \dots, p_k , show that the relation between the actual and the binomial variance of the overall proportion dying is

$$V_{\text{act}} = V_{\text{bin}} - \sum (p_i - p)^2/nk^2$$

EXAMPLE 16 12 3—In a sample of size n with population probability p the true logit is $\ln(p/q)$. The value $Y = \ln\{(g + \frac{1}{2})/(n - g + \frac{1}{2})\}$ is a relatively unbiased estimate of

$\ln(p/q)$ for expectations np and nq as low as 3 The weight $W = (g + \frac{1}{2})(n - g + \frac{1}{2})/(n + 1)$ corresponds to a variance

$$V = \frac{1}{W} = \frac{1}{g + \frac{1}{2}} + \frac{1}{n - g + \frac{1}{2}}$$

The quantity V is an almost unbiased estimate of the population variance of Y in small samples As an illustration the values of the binomial probability P and of Y and V are shown below for each value of g when $n = 10$ $p = 0.3$

g	P	Y	V	Y^2
0	0282	-3.046	2.095	9.278
1	1211	-1.846	0.772	3.408
2	2335	-1.224	0.518	1.498
3	2668	-0.762	0.419	0.581
4	2001	-0.367	0.376	0.135
5	1029	0.000	0.364	0.000
6	0368	0.367	0.376	0.135
7	0090	0.762	0.419	0.581
8	0014	1.224	0.518	1.498
9	0001	1.846	0.772	3.408
10	0000	3.046	2.095	9.278

The true logit is $\ln(0.3/0.7) = -0.8473$ Verify that (i) the mean value of Y is -0.8497 (ii) the variance of Y is 0.4968 (iii) the mean value of V is 0.5164 about 4%, too large

REFERENCES

- 1 J W GOWEN *Amer J Hum Genet* 4 285 (1952)
- 2 A E BRANDT Ph D Thesis Iowa State College (1932)
- 3 M B WILK and O KFMPTHORNE WADC Technical Report 55-244 Vol II Office of Technical Services U S Dept of Commerce Washington D C (1956)
- 4 N STRAND and R J JESSEN Iowa Agric Exp Sta Res Bul 315 (1943)
- 5 G W SNEDECOR and W R BRENNEMAN Iowa State College J Sci 19 333 (1945)
- 6 B BROWN *Proc Iowa Acad Sci* 38 205 (1932)
- 7 F YATES *J Amer Stat Ass* 29 51 (1934)
- 8 W L STEVENS *Biometrika* 35 346 (1948)
- 9 W T FEDERER and M ZILLEN *Biometrics* 22 525 (1966)
- 10 E R BECKER and P R HALL *Parasitology* 25 397 (1933)
- 11 W G COCHRAN *Ann Math Statist* 11 335 (1940)
- 12 M S BARTLETT *J R Statist Soc Supp* 2 248 (1935)
- 13 G A SILVER *Family Medical Care* Harvard University Press Cambridge Mass Table 59 (1963)
- 14 J I GART and J R ZWILLEN *Biometrika* 54 315 (1967)
- 15 H A KAHN Nat Canc Inst Monograph 19 1 (1966)
- 16 F YATES *Sampling Methods for Censuses and Surveys* Charles Griffin London 3rd ed Section 9.7 (1960)
- 17 G V DYKE and H D PATTERSON *Biometrics* 8 1 (1952)
- 18 L A GOODMAN *J Amer Statist Ass* 59 319 (1964)

Design and analysis of sampling

17.1—Populations. In the 1908 paper in which he discovered the *t*-test, “Student” opened with the following words: “Any experiment may be regarded as forming an individual of a *population* of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

“Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong.”

From the previous chapters in this book, this way of looking at data should now be familiar. The data obtained in an experiment are subject to variation, so that an estimate made from the data is also subject to variation and is, hence, to some degree uncertain. You can visualize, however, that if you could repeat the experiment many times, putting all the results together, the estimate would ultimately settle down to some unchanging value which may be called the true or definitive result of the experiment. The purpose of the statistical analysis of an experiment is to reveal what the data can tell about this true result. The tests of significance and confidence limits which have appeared throughout this book are tools for making statements about the population of experiments of which your data are a sample.

In such problems the sample is concrete, but the population may appear somewhat hypothetical. It is the population of experiments that might be performed, under the same conditions, if you possessed the necessary resources, time, and interest.

In this chapter we turn to situations in which the population is concrete and definite, and the problem is to obtain some desired information about it. Examples are as follows:

<i>Population</i>	<i>Information Wanted</i>
Ears of corn in a field	Average moisture content
Seeds in a large batch	Percentage germination
Water in a reservoir	Concentration of certain bacteria
Third-grade children in a school	Average weight

If the population is small, it is sometimes convenient to obtain the information by collecting the data for the whole of the population. More frequently, time and money can be saved by measuring only a sample drawn from the population. When the measurement is destructive, sampling is of course unavoidable.

This chapter presents some methods for selecting a sample and for estimating population characteristics from the data obtained in the sample. During the past thirty years, sampling has come to be relied upon by a great variety of agencies, including government bureaus, market research organizations, and public opinion polls. Concurrently, much has been learned both about the theory and practice of sampling, and a number of books devoted to sample survey methods have appeared (2, 3, 4, 5, 13). In this chapter we explain the general principles of sampling and show how to handle some of the simpler problems that are common in biological work. For more complex problems, references will be given.

17.2—A simple example. In the early chapters of this book, you drew samples so as to examine the amount of variation in results from one sample to another and to verify some important results in statistical theory. The same method will illustrate modern ideas about the selection of samples from given populations.

Suppose the population consists of $N = 6$ members, denoted by the letters a to f . The six values of the quantity that is being measured are as follows: a 1; b 2; c 4; d 6; e 7; f 16. The total for this population is 36. A sample of three members is to be drawn in order to estimate this total.

One procedure already familiar to you is to write the letters a to f on beans or slips of paper, mix them in some container, and draw out three letters. In sample survey work, this method of drawing is called *simple random sampling*, or sometimes *random sampling without replacement* (because we do not put a letter back in the receptacle after it has been drawn). Obviously, simple random sampling gives every member an equal chance of being in the sample. It may be shown that the method also gives every combination of three different letters (e.g., ae or cde) an equal chance of constituting the sample.

How good an estimate of the population total do we obtain by simple random sampling? We are not quite ready to answer this question. Although we know how the sample is to be drawn, we have not yet discussed how the population total is to be estimated from the results of the sample. Since the sample contains three members and the population contains six members, the simplest procedure is to multiply the sample total by 2, and this is the procedure that will be adopted. You should note that any sampling plan contains two parts—a rule for drawing the sample and a rule for making the estimates from the results of the sample.

We can now write down all possible samples of size 3, make the estimate from each sample, and see how close these estimates lie to the true value of 36. There are 20 different samples. Their results appear in table 17.2.1, where the successive columns show the composition of the sample,

the sample total, the estimated population total, and the error of estimate (estimate *minus* true value).

Some samples, e.g., *abf* and *cde*, do very well, while others like *abc* give poor estimates. Since we do not know in any individual instance whether we will be lucky or unlucky in the choice of a sample, we appraise any sampling plan by looking at its *average* performance.

TABLE 17.2.1
RESULTS FOR ALL POSSIBLE SIMPLE RANDOM SAMPLES OF SIZE THREE

Sample	Sample Total	Estimate of Population Total	Error of Estimate	Sample	Sample Total	Estimate of Population Total	Error of Estimate
<i>abc</i>	7	14	-22	<i>bcd</i>	12	24	-12
<i>abd</i>	9	18	-18	<i>bce</i>	13	26	-10
<i>abe</i>	10	20	-16	<i>bcf</i>	22	44	+ 8
<i>abf</i>	19	38	+ 2	<i>bde</i>	15	30	- 6
<i>acd</i>	11	22	-14	<i>bdf</i>	24	48	+12
<i>ace</i>	12	24	-12	<i>bef</i>	25	50	+14
<i>acf</i>	21	42	+ 6	<i>cde</i>	17	34	- 2
<i>ade</i>	14	28	- 8	<i>cdf</i>	26	52	+16
<i>adf</i>	23	46	+10	<i>cef</i>	27	54	+18
<i>aef</i>	24	48	+12	<i>def</i>	29	58	+22
				Average	18	36	0

The average of the errors of estimate, taking account of their signs, is called the *bias* of the estimate (or, more generally, of the sampling plan). A positive bias implies that the sampling plan gives estimates that are on the whole too high; a negative bias, too low. From table 17.2.1 it is evident that this plan gives unbiased estimates, since the average of the 20 estimates is exactly 36 and consequently the errors of estimate add to zero. With simple random sampling this result holds for any population and any size of sample. Estimates that are unbiased are a desirable feature of a sampling plan. On the other hand, a plan that gives a small bias is not ruled out of consideration if it has other attractive features.

As a measure of the accuracy of the sampling plan we use the mean square error of the estimates taken about the true population value. This is

$$M.S.E. = \frac{\Sigma(\text{Error of estimate})^2}{20} = \frac{3,504}{20} = 175.2$$

The divisor 20 is used instead of the divisor 19, because the errors are measured from the true population value. To sum up, this plan gives an estimate of the population total that is unbiased and has a standard error $\sqrt{175.2} = 13.2$. This standard error amounts to 37% of the true population total; evidently the plan is not very accurate for this population.

In simple random sampling the selection of the sample is left to the luck of the draw. No use is made of any knowledge that we possess about the members of the population. Given such knowledge, we should be able to improve upon simple random sampling by using the knowledge to guide us in the selection of the sample. Much of the research on sample survey methods has been directed towards taking advantage of available information about the population to be sampled.

By way of illustration, suppose that before planning the sample we expect that f will give a much higher value than any other member in the population. How can we use this information? It is clear that the estimate from the sample will depend to a considerable extent on whether f falls in the sample or not. This statement can be verified from table 17.2.1; every sample containing f gives an overestimate and every sample without f gives an underestimate.

The best plan is to make sure that f appears in every sample. We can do this by dividing the population into two parts or *strata*. Stratum I, which consists of f alone, is completely measured. In stratum II, containing a, b, c, d , and e , we take a simple random sample of size 2 in order to keep the total sample size equal to 3.

Some forethought is needed in deciding how to estimate the population total. To use twice the sample total, as was done previously, gives too much weight to f and, as already pointed out, will always produce an overestimate of the true total. We can handle this problem by treating the two strata separately. For stratum I we know the total (16) correctly, since we always measure f . For stratum II, where 2 members are measured out of 5, the natural procedure is to multiply the sample total in that stratum by 5/2, or 2.5. Hence the appropriate estimate of the population total is

$$16 + 2.5 (\text{Sample total in stratum II})$$

These estimates are shown for the 10 possible samples in table 17.2.2. Again we note that the estimate is unbiased. Its mean square error is

$$\frac{\Sigma (\text{Error of estimate})^2}{10} = \frac{487.50}{10} = 48.75$$

The standard error is 7.0 or 19% of the true total. This is a marked improvement over the standard error of 13.2 that was obtained with simple random sampling.

This sampling plan goes by the name of *stratified random sampling with unequal sampling fractions*. The last part of the title denotes the fact that stratum I is completely sampled, whereas stratum II is sampled at a rate of 2 units out of 5, or 40%. Stratification allows us to divide the population into sub-populations or strata that are less variable than the

TABLE 17.2.2
RESULTS FOR ALL POSSIBLE STRATIFIED RANDOM SAMPLES WITH THE UNEQUAL
SAMPLING FRACTIONS DESCRIBED IN TEXT

Sample	Sample Total in Stratum II (T_2)	Estimate $16 + 2.5 T_2$	Error of Estimate
<i>abf</i>	3	23.5	-12.5
<i>acf</i>	5	28.5	-7.5
<i>adf</i>	7	33.5	-2.5
<i>aef</i>	8	36.0	0.0
<i>bcf</i>	6	31.0	-5.0
<i>bdf</i>	8	36.0	0.0
<i>bef</i>	9	38.5	+2.5
<i>cdf</i>	10	41.0	+5.0
<i>cef</i>	11	43.5	+7.5
<i>def</i>	13	48.5	+12.5
Average		36.0	0.0

original population, and to sample different parts of the population at different rates when this seems advisable. It is discussed more fully in sections 17.8 and 17.9.

EXAMPLE 17.2.1—In the preceding example, suppose you expect that both *e* and *f* will give high values. You decide that the sample shall consist of *e*, *f*, and one member drawn at random from *a*, *b*, *c*, *d*. Show how to obtain an unbiased estimate of the population total and show that the standard error of this estimate is 7.7. (This sampling plan is not as accurate as the plan in which *f* alone was placed in a separate stratum, because the actual value for *e* is not very high.)

EXAMPLE 17.2.2—If previous information suggests that *f* will be high, *d* and *e* moderate, and *a*, *b*, and *c* small, we might try stratified sampling with three strata. The sample consists of *f*, either *d* or *e*, and one chosen from *a*, *b*, and *c*. Work out the unbiased estimate of the population total for each of the six possible samples and show that its standard error is 3.9.

17.3—Probability sampling. The preceding examples were intended to introduce you to *probability sampling*. This is a general name given to sampling plans in which

- (i) every member of the population has a known probability of being included in the sample,
- (ii) the sample is drawn by some method of random selection consistent with these probabilities,
- (iii) we take account of these probabilities of selection in making the estimates from the sample.

Note that the probability of selection need not be equal for all members of the population: it is sufficient that these probabilities be known. In the first example in the previous section, each member of the population had an equal chance of being in the sample, and each member of the sample received an equal weight in estimating the population total. But in the second example, member *f* was given a probability 1 of appearing in the sample, as against 2/5 for the rest of the population. This inequality in

the probabilities of selection was compensated for by assigning a weight $5/2$ to these other members when making the estimate. The use of unequal probabilities produces a substantial gain in precision for some types of populations (see section 17.9).

Probability sampling has several advantages. By probability theory it is possible to study the biases and the standard errors of the estimates from different sampling plans. In this way much has been learned about the scope, advantages, and limitations of each plan. This information helps greatly in selecting a suitable plan for a particular sampling job. As will be seen later, most probability sampling plans also enable the standard error of the estimate, and confidence limits for the true population value, to be computed from the results of the sample. Thus, when a probability sample has been taken, we have some idea as to how accurate the estimates are.

Probability sampling is by no means the only way of selecting a sample. An alternative method is to ask someone who has studied the population to point out "average" or "typical" members, and then confine the sample to these members. When the population is highly variable and the sample is small, this method often gives more accurate estimates than probability sampling. Another method is to restrict the sampling to those members that are conveniently accessible. If bales of goods are stacked tightly in a warehouse, it is difficult to get at the inside bales of the pile and one is tempted to confine attention to the outside bales. In many biological problems it is hard to see how a workable probability sample can be devised, as in estimating, for instance, the number of house flies in a town, or of field mice in a wood, or of plankton in the ocean.

One drawback of these alternative methods is that when the sample has been obtained, there is no way of knowing how accurate the estimate is. Members of the population picked out as typical by an expert may be more or less atypical. Outside bales may or may not be similar to interior bales. Probability sampling formulas for the standard error of the estimate or for confidence limits do not apply to these methods. Consequently, it is wise to use probability sampling unless there is a clear case that this is not feasible or is prohibitively expensive.

17.4—Listing the population. In order to apply probability sampling, we must have some way of subdividing the population into units, called *sampling units*, which form the basis for the selection of the sample. The sampling units must be distinct and non-overlapping, and they must together constitute the whole of the population. Further, in order to make some kind of random selection of sampling units, we must be able to number or *list* all the units. As will be seen, we need not always write down the complete list but we must be in a position to construct it. Listing is easily accomplished when the population consists of 5,000 cards neatly arranged in a file, or 300 ears of corn lying on a bench, or the trees in a small orchard. But the subdivision of a population into sampling units that can be listed sometimes presents a difficult practical problem.

Although we have spoken of the population as being concrete and definite, there may be some vagueness about the population which does not become apparent until a sampling is being planned. Before we can come to grips with a population of farms or of nursing homes, we must define a farm or a nursing home. The definition may require much study and the final decision may have to be partly arbitrary. Two principles to keep in mind are that the definition should be appropriate to the purpose of the sampling and that it should be usable in the field (i.e., the person collecting the information should be able to tell what is in and what is out of the population as defined).

Sometimes the available listings of farms, creameries, or nursing homes are deficient. The list may be out of date, having some members that no longer belong to our population and omitting some that do belong. The list may be based on a definition different from that which we wish to use for our population. These points should be carefully checked before using any list. It often pays to spend considerable effort in revising a list to make it complete and satisfactory, since this may be more economical than constructing a new list. Where a list covers only part of the population, one procedure is to sample this part by means of the list, and to construct a separate method of sampling for the unlisted part of the population. Stratified sampling is useful in this situation: all listed members are assigned to one stratum and unlisted members to another.

Preparing a list where none is available may require ingenuity and hard work. To cite an easy example, suppose that we wish to take a number of crop samples, each 2 ft. \times 2 ft., from a plot 200 ft. \times 100 ft. Divide the length of the plot into 100 sections, each 2 ft., and the breadth into 50 sections, each 2 ft. We thus set up a coordinate system that divides the whole plot into 100 \times 50 or 5,000 quadrats, each 2 ft. \times 2 ft. To select a quadrat by simple random sampling, we draw a random number between 1 and 100 and another random number between 1 and 50. These coordinates locate the corner of the quadrat that is farthest from the origin of our system. However, the problem becomes harder if the plot measures 163 ft. \times 100 ft., and much harder if we have an irregularly shaped field. Further, if we have to select a number of areas each 6 in. \times 6 in. from a large field, giving every area an equal chance of selection, the time spent in selecting and locating the sample areas becomes substantial. Partly for this reason, methods of systematic sampling (section 17.7) have come to be favored in routine soil sampling (8).

Another illustration is a method for sampling (for botanical or chemical analysis) the produce of a small plot that is already cut and bulked. The bulk is separated into two parts and a coin is tossed (or a random number drawn) to decide which part shall contain the sample. This part is then separated into two, and the process continues until a sample of about the desired size is obtained. At any stage it is good practice to make the two parts as alike as possible, provided this is done before the coin is tossed. A quicker method, of course, is to grab a handful of about the

desired size; this is sometimes satisfactory but sometimes proves to be biased.

In urban sampling in the United States, the city block is often used as a sampling unit, a listing of the blocks being made from a map of the town. For extensive rural sampling, county maps have been divided into areas with boundaries that can be identified in the field and certain of these areas are selected to constitute the sample. The name *area sampling* has come to be associated with these and other methods in which the sampling unit is an area of land. Frequently the principal advantage of area sampling, although not the only one, is that it solves the problem of providing a listing of the population by sampling units.

In many sampling problems there is more than one type or size of sampling unit into which the population can be divided. For instance, in soil sampling in which borings are taken, the size and shape of the borer can be chosen by the sampler. The same is true of the frame used to mark out the area of land that is cut in crop sampling. In a dental survey of the fifth-grade school children in a city, we might regard the child as the sampling unit and select a sample of children from the combined school registers for the city. It would be administratively simpler, however, to take the school as the sampling unit, drawing a sample of schools and examining every fifth-grade child in the selected schools. This approach, in which the sampling unit consists of some natural group (the school) formed from the smaller units in which we are interested (the children), goes by the name of *cluster sampling*.

If you are faced with a choice between different sampling units, the guiding rule is to try to select the one that returns the greatest precision for the available resources. For a fixed size of sample (e.g., 5% of the population), a large sampling unit usually gives less accurate results than a small unit, although there are exceptions. To counterbalance this, it is generally cheaper and easier to take a 5% sample with a large sampling unit than with a small one. A thorough comparison between two units is likely to require a special investigation, in which both sampling errors and costs (or times required) are computed for each unit.

17.5—Simple random sampling. In this and later sections, some of the best-known methods for selecting a probability sample will be presented. The goal is to use a sampling plan that gives the highest precision for the resources to be expended, or, equivalently, that attains a desired degree of precision with the minimum expenditure of resources. It is worthwhile to become familiar with the principal plans, since they are designed to take advantage of any information that you have about the structure of the population and about the costs of taking the sample.

In section 17.2 you have already been introduced to *simple random sampling*. This is a method in which the members of the sample are drawn independently with equal probabilities. In order to illustrate the use of a table of random numbers for drawing a random sample, suppose that the population contains $N = 372$ members and that a sample of size $n = 10$

is wanted. Select a three-digit starting number from table A 1, say the number is 539 in row 11 of columns 80–82. Read down the column and pick out the first ten three-digit numbers that do not exceed 372. These are 334, 365, 222, 345, 245, 272, 075, 038, 127, and 112. The sample consists of the sampling units that carry these numbers in your listing of the population. If any number appears more than once, ignore it on subsequent appearances and proceed until ten *different* numbers have been found.

If the first digit in N is 1, 2, or 3, this method requires you to skip many numbers in the table because they are too large. (In the above example we had to cover 27 numbers in order to find ten for the sample.) This does not matter if there are plenty of random numbers. An alternative is to use all three-digit numbers up to $2 \times 372 = 744$. Starting at the same place, the first ten numbers that do not exceed 744 are 539, 334, 615, 736, 365, 222, 345, 660, 431, and 427. Now subtract 372 from all numbers larger than 372. This gives, for the sample, 167, 334, 243, 364, 365, 222, 345, 288, 59, and 55. With $N = 189$, for instance, we can use all numbers up to $5 \times 189 = 945$ by this device, subtracting 189 or 378 or 567 or 756 as the case may be.

As mentioned previously, simple random sampling leaves the selection of the sample entirely to chance. It is often a satisfactory method where the population is not highly variable and, in particular, when estimating proportions that are likely to lie between 20% and 80%. On the other hand, if you have any knowledge of the variability in the population, such as that certain segments of it are likely to give higher responses than others, one of the methods to be described later may be more precise.

If Y_i ($i = 1, 2, \dots, N$) denotes the variable that is being studied, the standard deviation, σ , of the population is defined as

$$\sigma = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{N - 1}},$$

where \bar{Y} is the population mean of the Y_i and the sum Σ is taken over all sampling units in the population.

Since \bar{Y} denotes the population mean, we shall use \bar{y} to denote the sample mean. In a simple random sample of size n , the standard error of \bar{y} is:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} \sqrt{(1 - \phi)},$$

where $\phi = n/N$ is the *sampling fraction*, i.e., the fraction of the population that is included in the sample. The sampling fraction is commonly denoted by the symbol f , but ϕ is used here to avoid confusion with our previous use of f for degrees of freedom.)

The term σ/\sqrt{n} is already familiar to you: this is the usual formula for the standard error of a sample mean. The second factor, $\sqrt{(1 - \phi)}$, is

known as the *finite population correction*. It enters because we are sampling from a population of finite size, N , instead of from an infinite population as is assumed in the usual theory. Note that this term makes the standard error zero when $n = N$, as it should do, since we have then measured every unit in the population. In practical applications the finite population correction is close to 1 and can be omitted when n/N is less than 10%, i.e., when the sample includes less than 10% of the population.

This result is remarkable. In a large population with a fixed amount of variability (a given value of σ), the standard error of the mean depends mainly on the size of sample and only to a minor extent on the fraction of the population that is sampled. For given σ , the mean of a sample of 100 is almost as precise when the population size is 200,000 as when the population size is 20,000 or 2,000. Intuitively, some people feel that one cannot possibly get accurate results from a sample of 100 out of a population of 200,000, because only a tiny fraction of the population has been measured. Actually, whether the sampling plan is accurate or not depends primarily on the size of σ/\sqrt{n} . This shows why sampling can bring about a great reduction in the amount of measurement needed.

For the *estimated* standard error of the sample mean we have

$$s_y = \frac{s}{\sqrt{n}} \sqrt{(1 - \phi)},$$

where s is the standard deviation of the sample, calculated in the usual way.

If the sample is used to estimate the population *total* of the variable under study, the estimate is $N\bar{y}$ and its estimated standard error is

$$s_{N\bar{y}} = \frac{Ns}{\sqrt{n}} \sqrt{(1 - \phi)}$$

In simple random sampling for attributes, where every member of the sample is classified into one of two classes, we take

$$s_y = \sqrt{\frac{pq}{n}} \sqrt{(1 - \phi)},$$

where p is the proportion of the sample that lies in one of the classes. Suppose that 50 families are picked at random from a list of 432 families who possess telephones and that 10 of the families report that they are listening to a certain radio program. Then $p = 0.2$, $q = 0.8$ and

$$s_p = \sqrt{\frac{(0.2)(0.8)}{50}} \sqrt{\left(1 - \frac{50}{432}\right)} = 0.053$$

If we ignore the finite population correction, we find $s_p = 0.057$.

The formula for s_p holds only if each sampling unit is classified as a whole into one of the two classes. If you are using cluster sampling and are classifying individual elements within each cluster, a different formula for

s_p must be used. For instance, in estimating the percentage of diseased plants in a field from a sample of 360 plants, the formula above holds if the plants were selected independently and at random. To save time in the field, however, we might have chosen 40 areas, each consisting of 3 plants in each of 3 neighboring rows. With this method the area (a cluster of 9 plants) is the sampling unit. If the distribution of disease in the field were extremely patchy, it might happen that every area had either all 9 plants diseased or no plants diseased. In this event the sample of 40 areas would be no more precise than a sample of 40 independently chosen plants, and we would be deceiving ourselves badly if we thought that we had a binomial sample of 360 plants.

The correct procedure for computing s_p is simple. Calculate p separately for each area (or sampling unit) and apply to these p 's the previous formula for continuous variates. That is, if p_i is the percentage diseased in the i th area, the sample standard deviation is

$$s = \sqrt{\frac{\sum (p_i - p)^2}{(n - 1)}},$$

where n is now the number of areas (cluster units). Then

$$s_p = \frac{s}{\sqrt{n}} \sqrt{(1 - p)}$$

For instance, suppose that the numbers of diseased plants in the 40 areas were as given in table 17.5.1.

TABLE 17.5.1
NUMBERS OF DISEASED PLANTS (OUT OF 9) IN EACH OF 40 AREAS

2	5	1	1	1	7	0	0	3	2	3	0	0	0	7	0	4	1	2	6
0	0	1	4	5	0	1	4	2	6	0	2	4	1	7	3	5	0	3	6
Grand total = 99																			

The standard deviation of the numbers of diseased plants in this sample is 2.331. Since the *proportions* of diseased plants in the 40 areas are found by dividing the numbers in table 17.5.1 by 9, the standard deviation of the proportions is

$$s = \frac{2.331}{9} = 0.259$$

Hence (assuming N large),

$$s_p = \frac{s}{\sqrt{n}} = \frac{0.259}{\sqrt{40}} = 0.041$$

For comparison, the result given by the binomial formula will be worked out. From the total in table 17.5.1, $p = 99/360 = 0.275$. The binomial formula is

$$s_p = \sqrt{\frac{pq}{360}} = \sqrt{\frac{(0.275)(0.725)}{360}} = 0.024,$$

giving an overly optimistic notion of the precision of p .

Frequently, the clusters are not all of the same size. This happens when the sampling units are areas of land that contain different numbers of the plants that are being classified. Let m_i be the number of elements that are classified in the i th unit, and a_i the number that fall into a specified class, so that $p_i = a_i/m_i$. Then p , the overall proportion in the sample is $(\Sigma a_i)/(\Sigma m_i)$, where each sum is taken over the n cluster units.

The formula for s , the standard deviation of the individual proportions p_i uses a weighted mean square of the deviations $(p_i - p)$, as follows:

$$s = \sqrt{\frac{1}{(n-1)} \sum \left\{ \left(\frac{m_i}{\bar{m}} \right)^2 (p_i - p)^2 \right\}}$$

where $\bar{m} = \Sigma m_i/n$ is the average size of cluster in the sample. This formula is an approximation, no correct expression for s being known in usable form. As before, we have

$$s_p = \frac{s}{\sqrt{n}} \sqrt{(1 - \phi)}$$

For computing purposes, s is better expressed as

$$s = \frac{1}{\bar{m}} \sqrt{\frac{1}{(n-1)} \{ \Sigma a_i^2 - 2p \Sigma a_i m_i + p^2 \Sigma m_i^2 \}}$$

The sums of squares Σa_i^2 , Σm_i^2 and the sum of products $\Sigma a_i m_i$ are calculated without the usual corrections for the mean. The same value of s is obtained whether the corrections for the mean are applied or not, but it saves time not to apply them.

EXAMPLE 17.5.1—If a sample of 4 from the 16 townships of a county has a standard deviation 45, show that the standard error of the mean is 19.5

EXAMPLE 17.5.2—In the example presented in section 17.2 we had $N = 6$, $n = 3$, and the values for the 6 members of the population were 1, 2, 4, 6, 7, and 16. The formula for the true standard error of the estimated population total is

$$\sigma_{N\bar{y}} = \frac{N\sigma}{\sqrt{n}} \sqrt{\left(1 - \frac{n}{N}\right)}$$

Verify that this formula agrees with the result, 13.2, which we found by writing down all possible samples

EXAMPLE 17.5.3—A simple random sample of size 100 is taken in order to estimate some proportion (e.g., the proportion of males) whose value in the population is close to $1/2$. Work out the standard error of the sample proportion p when the size of the population is (i) 200, (ii) 500, (iii) 1,000, (iv) 10,000, (v) 100,000. Note how little the standard error changes for N greater than 1,000.

EXAMPLE 17.5.4—Show that the coefficient of variation of the sample mean is the same as that of the estimated population total.

EXAMPLE 17.5.5—In simple random sampling for attributes, show that the standard error of p , for given N and n , is greatest when p is 50%, but that the coefficient of variation of p is largest when p is very small.

17.6—Size of sample. At an early stage in the design of a sample, the question “How large a sample do I need?” must be considered. Although a precise answer may not be easy to find, for reasons that will appear, there is a rational method of attack on the problem.

Clearly, we want to avoid making the sample so small that the estimate is too inaccurate to be useful. Equally, we want to avoid taking a sample that is too large, in that the estimate is more accurate than we require. Consequently, the first step is to decide how large an error we can tolerate in the estimate. This demands careful thinking about the use to be made of the estimate and about the consequences of a sizeable error. The figure finally reached may be to some extent arbitrary, yet after some thought samplers often find themselves less hesitant about naming a figure than they expected to be.

The next step is to express the allowable error in terms of confidence limits. Suppose that L is the allowable error in the sample mean, and that we are willing to take a 5% chance that the error will exceed L . In other words, we want to be reasonably certain that the error will not exceed L . Remembering that the 95% confidence limits computed from a sample mean, assumed approximately normally distributed, are

$$\bar{y} \pm \frac{2\sigma}{\sqrt{n}},$$

where we have ignored the finite population correction, we put

$$L = \frac{2\sigma}{\sqrt{n}}$$

This gives, for the required sample size,

$$n = \frac{4\sigma^2}{L^2}$$

In order to use this relation, we must have an estimate of the population standard deviation, σ . Often a good guess can be made from the results of previous samplings of this population or of other similar populations. For example, an experimental sample was taken in 1938 to estimate

the yield per acre of wheat in certain districts of North Dakota (7). For a sample of 222 fields, the variance of the yield per acre from field to field was $s^2 = 90.3$ (in bushels²). How many fields are indicated if we wish to estimate the true mean yield within ± 1 bushel, with a 5% risk that the error will exceed 1 bushel? Then

$$n = \frac{4\sigma^2}{L^2} = \frac{4(90.3)}{(1)^2} = 361 \text{ fields}$$

If this estimate were being used to plan a sample in some later year, it would be regarded as tentative, since the variance between fields might change from year to year.

In default of previous estimates, Deming (3) has pointed out that σ can be estimated from a knowledge of the highest and lowest values in the population and a rough idea of the shape of the distribution. If h = (highest - lowest); then $\sigma = 0.29h$ for a uniform (rectangular) distribution, $\sigma = 0.24h$ for a symmetrical distribution shaped like an isosceles triangle, and $\sigma = 0.21h$ for a skew distribution shaped like a right triangle.

If the quantity to be estimated is a binomial proportion, the allowable error, L , for 95% confidence probability is

$$L = 2 \sqrt{\frac{pq}{n}}$$

The sample size required to attain a given limit of error, L , is therefore

$$n = \frac{4pq}{L^2} \quad (17.6.1)$$

In this formula, p , q , and L may be expressed either as proportions or as percentages, provided they are all expressed in the same units. The result necessitates an advance estimate of p . If p is likely to lie between 35% and 65%, the advance estimate can be quite rough, since the product pq varies little for p lying between these limits. If, however, p is near zero or 100%, accurate determination of n requires a close guess about the value of p .

We have ignored the finite population correction in the formulas presented in this section. This is satisfactory for the majority of applications. If the computed value of n is found to be more than 10% of the population size, N , a revised value n' which takes proper account of the correction is obtained from the relation

$$n' = \frac{n}{1 + \phi}$$

For example, casual inspection of a batch of 480 seedlings indicates that about 15% are diseased. Suppose we wish to know the size of sample needed to determine p , the per cent diseased, to within $\pm 5\%$, apart from a 1-in-20 chance. Formula 17.6.1 gives

$$n = \frac{4(15)(85)}{(25)} = 204 \text{ seedlings}$$

At this point we might decide that it would be as quick to classify every seedling as to plan a sample that is a substantial part of the whole batch. If we decide on sampling, we make a revised estimate, n' , as

$$n' = \frac{n}{1 + \phi} = \frac{204}{1 + \frac{204}{480}} = 143$$

The formulas presented in this section are appropriate for simple random sampling. If some other sampling method is to be used, the general principles for the determination of n remain the same, but the formula for the confidence limits, and hence the formula connecting L with n , will change. Formulas applicable to more complex methods of sampling can be obtained in books devoted to the subject, e.g., (2, 4). In practice, the formulas in this section are frequently used to provide a preliminary notion of the value of n , even if simple random sampling is not intended to be used. The values of n are revised later if the proposed method of sampling is markedly different in precision from simple random sampling.

When more than one variable is to be studied, the value of n is first estimated separately for each of the most important variables. If these values do not differ by much, it may be feasible to use the largest of the n 's. If the n 's differ greatly, one method is to use the largest n , but to measure certain items on only a sub-sample of the original sample, e.g., on 200 sampling units out of 1,000. In other situations, great disparity in the n 's is an indication that the investigation must be split into two or more separate surveys.

EXAMPLE 17.6.1—A simple random sample of houses is to be taken to estimate the percentage of houses that are unoccupied. The estimate is desired to be correct to within $\pm 1\%$, with 95% confidence. One advance estimate is that the percentage of unoccupied houses will be about 6%, another is that it will be about 4%. What sizes of sample are required on these two forecasts? What size would you recommend?

EXAMPLE 17.6.2 The total number of rats in the residential part of a large city is to be estimated with an error of not more than 20%, apart from a 1-in-20 chance. In a previous survey, the mean number of rats per city block was nine and the sample standard deviation was 19 (the distribution is extremely skew). Show that a simple random sample of around 450 blocks should suffice.

EXAMPLE 17.6.3 West (12) quotes the following data for 556 full-time farms in Seneca County, New York.

	Mean	Standard Deviation Per Farm
Acres in corn	8.8	9.0
Acres in small grains	42.0	39.5
Acres in hay	27.9	26.9

If a coefficient of variation of up to 5% can be tolerated, show that a random sample of about 240 farms is required to estimate the total acreage of each crop in the 556 farms with this degree of precision (Note that the finite population correction must be used) This example illustrates a result that has been reached by several different investigators, with small farm populations such as counties, a substantial part of the whole population must be sampled in order to obtain accurate estimates

17.7—Systematic sampling. In order to draw a 10% sample from a list of 730 cards, we might select a random number between 1 and 10, say 3, and pick every 10th card thereafter; i.e., the cards numbered 3, 13, 23, and so on, ending with the card numbered 723. A sample of this kind is known as a *systematic sample*, since the choice of its first member, 3, determines the whole sample.

Systematic sampling has two advantages over simple random sampling. It is easier to draw, since only one random number is required, and it distributes the sample more evenly over the listed population. For this reason systematic sampling often gives more accurate results than simple random sampling. Sometimes the increase in accuracy is large. In routine sampling, systematic selection has become a popular technique.

There are two potential disadvantages. If the population contains a periodic type of variation, and if the interval between successive units in the systematic sample happens to coincide with the wave length (or a multiple of it) we may obtain a sample that is badly biased. To cite extreme instances, a systematic sample of the houses in a city might contain far too many, or too few, corner houses; a systematic sample from a book of names might contain too many, or too few, names listed first on a page, who might be predominantly males, or heads of households, or persons of importance. A systematic sample of the plants in a field might have the selected plants at the same positions along every row. These situations can be avoided by being on the lookout for them and either using some other method of sampling or selecting a new random number frequently. In field sampling, we could select a new random number in each row. Consequently, it is well to know something about the nature of the variability in the population before deciding to use systematic sampling.

The second disadvantage is that from the results of a systematic sample there is no reliable method of estimating the standard error of the sample mean. Textbooks on sampling give various formulas for s , that may be tried: each formula is valid for a certain type of population, but a formula can be used with confidence only if we have evidence that the population is of the type to which the formula applies. However, systematic sampling often is a part of a more complex sampling plan in which it is possible to obtain unbiased estimates of the sampling errors.

EXAMPLE 17.7.1 -The purpose of this example is to compare simple random sampling and systematic sampling of a small population. The following data are the weights of maize (in 10-gm. units) for 40 successive hills lying in a single row: 104, 38, 105, 86, 63, 32, 47, 0, 80, 42, 37, 48, 85, 66, 110, 0, 73, 65, 101, 47, 0, 36, 16, 33, 22, 32, 33, 0, 35, 82, 37, 45, 30

76, 45, 70, 70, 63, 83, 34. To save you time, the population standard deviation is given as 30.1. Compute the standard deviation of the mean of a simple random sample of 4 hills. A systematic sample of 4 hills can be taken by choosing a random number between 1 and 10 and taking every 10th hill thereafter. Find the mean \bar{y}_{sy} for each of the 10 possible systematic samples and compute the standard deviation of these means about the true mean \bar{Y} of the population. Note that the formula for the standard deviation is

$$\sigma(\bar{y}_{sy}) = \sqrt{\frac{\sum(\bar{y}_{sy} - \bar{Y})^2}{10}}$$

Verify that the standard deviation of the estimate is about 8% lower with systematic sampling. To what do you think this difference is due?

17.8—Stratified sampling. There are three steps in stratified sampling:

- (1) The population is divided into a number of parts, called *strata*.
- (2) A sample is drawn independently in each part.
- (3) As an estimate of the population mean, we use

$$\bar{y}_{st} = \frac{\sum N_h \bar{y}_h}{N},$$

where N_h is the total number of sampling units in the h th stratum, \bar{y}_h is the sample mean in the h th stratum and $N = \sum N_h$ is the size of the population. Note that we must know the values of the N_h (i.e., the sizes of the strata) in order to compute this estimate.

Stratification is commonly employed in sampling plans for several reasons. It can be shown that differences between the strata means in the population do not contribute to the sampling error of the estimate \bar{y}_{st} . In other words, the sampling error of \bar{y}_{st} arises solely from variations among sampling units that are in the same stratum. If we can form strata so that a heterogeneous population is divided into parts each of which is fairly homogeneous, we may expect a gain in precision over simple random sampling. In taking 24 soil or crop samples from a rectangular field, we might divide the field into 12 compact plots, and draw 2 samples at random from each plot. Since a small piece of land is usually more homogeneous than a large piece, this stratification will probably bring about an increase in precision, although experience indicates that in this application the increase will be modest rather than spectacular. To estimate total wheat acreage from a sample of farms, we might stratify by size of farm, using any information available for this purpose. In this type of application the gain in precision is frequently large.

In stratified sampling, we can choose the size of sample that is to be taken from any stratum. This freedom of choice gives us scope to do an efficient job of allocating resources to the sampling within strata. In some applications, this is the principal reason for the gain in precision from stratification. Further, when different parts of the population present different problems of listing and sampling, stratification enables these

problems to be handled separately. For this reason, hotels and large apartment houses are frequently placed in a separate stratum in a sample of the inhabitants of a city.

We now consider the estimate from stratified sampling and its standard error. For the population mean, the estimate given previously may be written

$$\bar{y}_{st} = \frac{1}{N} \sum N_h \bar{y}_h = \sum W_h \bar{y}_h,$$

where $W_h = N_h/N$ is the relative *weight* attached to the stratum. Note that the sample means, \bar{y}_h , in the respective strata are weighted by the sizes, N_h , of the strata. The arithmetic mean of the sample observations is no longer the estimate except in one important special case. This occurs with *proportional allocation*, when we sample the same fraction from every stratum. With proportional allocation,

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \dots = \frac{n_h}{N_h} = \frac{n}{N}$$

It follows that

$$W_h = \frac{N_h}{N} = \frac{n_h}{n}$$

Hence,

$$\bar{y}_{st} = \sum W_h \bar{y}_h = \frac{\sum n_h \bar{y}_h}{n} = \bar{y},$$

since $\sum n_h \bar{y}_h$ is the total of all observations in the sample. With proportional allocation, we are saved the trouble of computing a weighted mean: the sample is *self-weighting*.

In order to avoid two levels of subscripts, we use the symbol $s(\bar{y}_{st})$ to denote the estimated standard error of \bar{y}_{st} . Its value is

$$s(\bar{y}_{st}) = \sqrt{\sum W_h^2 \frac{s_h^2}{n_h}},$$

where s_h^2 is the sample variance in the h th stratum, i.e.,

$$s_h^2 = \frac{\sum (Y_{hi} - \bar{y}_h)^2}{n_h - 1},$$

where Y_{hi} is the i th member of the sample from the h th stratum. This formula for the standard error of \bar{y}_{st} assumes that simple random sampling is used within each stratum and does not include the finite population

correction. If the sampling fractions ϕ_h exceed 10% in some of the strata, we use the more general formula

$$s(\bar{y}_{st}) = \sqrt{\sum \frac{W_h^2 s_h^2}{n_h} (1 - \phi_h)} \tag{17.8.1}$$

With proportional allocation the sampling fractions ϕ_h are all equal and the general formula simplifies to

$$s(\bar{y}_{st}) = \sqrt{\frac{\sum W_h s_h^2}{n}} \cdot \sqrt{(1 - \phi)}$$

If, further, the population variances are the same in all strata (a reasonable assumption in some applications), we obtain an additional simplification to

$$s(\bar{y}_{st}) = \frac{s_w}{\sqrt{n}} \sqrt{(1 - \phi)}$$

This result is the same as that for the standard error of the mean with simple random sampling, except that s_w , the pooled standard deviation *within strata*, appears in place of the sample standard deviation, s . In practice, s_w is computed from an analysis of variance of the data.

As an example of proportional allocation, the data in table 17.8.1 come from an early investigation by Clapham (1) of the feasibility of sampling for estimating the yields of small cereal plots. A rectangular plot of wheat was divided transversely into three equal strata. Ten samples, each a meter length of a single row, were chosen by simple random sampling from each stratum. The problem is to compute the standard error of the estimated mean yield per meter of row.

TABLE 17.8.1
ANALYSIS OF VARIANCE OF A STRATIFIED RANDOM SAMPLE
(Wheat grain yields — gm. per meter)

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Total	29	8,564	295.3
Between strata	2	2,073	1,036.5
Within strata	27	6,491	240.4

In this example, $s_w = \sqrt{240.4} = 15.5$, and $n = 30$. Since the sample is only a negligible part of the whole plot, n/N is negligible and

$$s(\bar{y}_{st}) = \frac{s_w}{\sqrt{n}} = \frac{15.5}{\sqrt{30}} = 2.83 \text{ gm.}$$

How effective was the stratification? From the analysis of variance it is seen that the mean square between strata is over four times as large as that within strata. This is an indication of real differences in level of yield from stratum to stratum. It is possible to go further, and estimate what the standard error of the mean would have been if simple random sampling had been used without any stratification. With simple random sampling, the corresponding formula for the standard error of the mean is

$$s_{\bar{y}} = \frac{s}{\sqrt{n}},$$

where s is the ordinary sample standard deviation. In the sample under discussion, s is $\sqrt{295.3}$ (from the *total* mean square in table 17.8.1). Hence, as an estimate of the standard error of the mean under simple random sampling, we might take

$$s_{\bar{y}} = \frac{\sqrt{295.3}}{\sqrt{30}} = 3.14 \text{ gm.},$$

as compared with 2.83 gm. for stratified random sampling. Stratification has reduced the standard error by about 10%.

This comparison is not quite correct, for the rather subtle reason that the value of s was calculated from the results of a stratified sample and not, as it should have been, from the results of a simple random sample. Valid methods of making the comparison are described for all types of stratified sampling in (2). The approximate method which we used is close enough when the stratification is proportional and at least ten sampling units are drawn from every stratum.

EXAMPLE 17.8.1—In the example of stratified sampling given in section 17.2, show that the estimate which we used for the population total was $N\bar{y}_w$. From formula 17.8.1 for the standard error of \bar{y}_w , verify that the variance of the estimated population total is 48.75, as found directly in section 17.2. (Note that stratum I makes no contribution to this variance because $n_h = N_h$ in that stratum.)

17.9—Choice of sample sizes in the individual strata. It is sometimes thought that in stratified sampling we should sample the same fraction from every stratum; i.e., we should make n_h/N_h the same in all strata, using proportional allocation. A more thorough analysis of the problem shows, however, that the *optimum* allocation is to take n_h proportional to $N_h\sigma_h/\sqrt{c_h}$, where σ_h is the standard deviation of the sampling units in the h th stratum, and c_h is the cost of sampling per unit in the h th stratum. This method of allocation gives the smallest standard error of the estimated mean \bar{y}_{st} for a given total cost of taking the sample. The rule tells us to take a larger sample, as compared with proportional allocation, in a stratum that is unusually variable (σ_h large), and a smaller sample in a stratum where sampling is unusually expensive (c_h large). Looked at in

this way, the rule is consistent with common sense, as statistical rules always are if we think about them carefully. The rule reduces to proportional allocation when the standard deviation and the cost per unit are the same in all strata.

In order to apply the rule, advance estimates are needed both of the relative standard deviations and of the relative costs in different strata. These estimates need not be highly accurate; rough estimates often give results satisfactorily near to the optimum allocation. When a population is sampled repeatedly, the estimates can be obtained from the results of previous samplings. Even when a population is sampled for the first time, it is sometimes obvious that some strata are more accessible to sampling than others. In this event it pays to hazard a guess about the differences in costs. In other situations we are unable to predict with any confidence which strata will be more variable or more costly, or we think that any such differences will be small. Proportional allocation is then used.

There is one common situation in which disproportionate sampling pays large dividends. This occurs when the principal variable that is being measured has a highly skewed or asymmetrical distribution. Usually, such populations contain a few sampling units that have large values for this variable and many units that have small values. Variables that are related to the sizes of economic institutions are often of this type, for instance, the total sales of grocery stores, the number of patients per hospital, the amounts of butter produced by creameries, family incomes, and prices of houses.

With populations of this type, stratification by size of institution is highly effective, and the optimum allocation is likely to be much better than proportional allocation. As an illustration, table 17.9.1 shows data for the number of students per institution in a population consisting of the 1,019 senior colleges and universities in the United States. The data, which apply mostly to the 1952-1953 academic year, might be used

TABLE 17.9.1
DATA FOR TOTAL REGISTRATIONS PER SENIOR COLLEGE OR UNIVERSITY,
ARRANGED IN FOUR STRATA

Stratum: Number of Students Per Institution	Number of Institutions N_h	Total Registration for the Stratum	Mean Per Institution \bar{Y}_h	Standard Deviation Per Institution σ_h
Less than 1,000	661	292,671	443	236
1,000-3,000	205	345,302	1,684	625
3,000-10,000	122	672,728	5,514	2,008
Over 10,000	31	573,693	18,506	10,023
Total	1,019	1,884,394		

as background information for planning a sample designed to give a quick estimate of total registration in some future year. The institutions are arranged in four strata according to size.

Note that the 31 largest universities, about 3% in number, have 30% of the students, while the smallest group, which contains 65% of the institutions, contributes only 15% of the students. Note also that the within-stratum standard deviation, σ_h , increases rapidly with increasing size of institution.

Table 17.9.2 shows the calculations needed for choosing the optimum sample sizes within strata. We are assuming equal costs per unit within all strata. The products, $N_h\sigma_h$, are formed and added over all strata. Then the relative sample sizes, $N_h\sigma_h/\Sigma N_h\sigma_h$, are computed. These ratios, when multiplied by the intended sample size n , give the sample sizes in the individual strata.

TABLE 17.9.2
CALCULATIONS FOR OBTAINING THE OPTIMUM SAMPLE SIZES IN INDIVIDUAL STRATA

Stratum: Number of Students	Number of Institutions N_h	$N_h\sigma_h$	Relative Sample Sizes $N_h\sigma_h/\Sigma N_h\sigma_h$	Actual Sample Sizes	Sampling Rate (%)
Less than 1,000	661	155,996	.1857	65	10
1,000-3,000	205	128,125	.1526	53	26
3,000-10,000	122	244,976	.2917	101	83
Over 10,000	31	310,713	.3700	31	100
Total	1,019	839,810	1.0000	250	

As a consequence of the large standard deviation in the stratum with the largest universities, the rule requires 37% of the sample to be taken from this stratum. Suppose we are aiming at a total sample size of 250. The rule then calls for $(0.37)(250)$ or 92 universities from this stratum although the stratum contains only 31 universities in all. With highly skewed populations, as here, the optimum allocation may demand 100% sampling, or even more than this, of the largest institutions. When this situation occurs, the best procedure is to take 100% of the "large" stratum, and employ the rule to distribute the remainder of the sample over the other strata. Following this procedure, we include in the sample all 31 largest institutions, leaving 219 to be distributed among the first three strata. In the first stratum, the size of sample is

$$219 \left\{ \frac{0.1857}{0.1857 + 0.1526 + 0.2917} \right\} = 65$$

The allocations, shown in the second column from the right of table 17.9.2, call for over 80% sampling in the second largest group of institu-

tions (101 out of 122), but only a 10% sample of the small colleges. In practice we might decide, for administrative convenience, to take a 100% sample in the second largest group as well as in the largest.

It is worthwhile to ask: Is the optimum allocation much superior to proportional allocation? If not, there is little value in going to the extra trouble of calculating and using the optimum allocation. We cannot, of course, answer this question for a future sample that is not yet taken, but we can compare the two methods of allocation for the 1952–1953 registrations. To do this, we use the data in tables 17.9.1 and 17.9.2 and the standard error formulas in section 17.8 to compute the standard errors of the estimated population totals by the two methods. These standard errors are found to be 26,000 for the optimum allocation, as against 107,000 for proportional allocation. If simple random sampling had been used, with no stratification, a similar calculation shows that the corresponding standard error would have been 216,000. The reduction in the standard error due to stratification, and the additional reduction due to the optimum allocation, are both striking. In an actual future sampling based on this stratification, the gains in precision would presumably be slightly less than these figures indicate.

EXAMPLE 17.9.1—For the population of colleges and universities discussed in this section it was stated that a stratified sample of 250 institutions, with proportional allocation, would have a standard error of 107,000 for the estimated total registration in all 1,019 institutions. Verify this statement from the data in table 17.9.1. Note that the standard error of the estimated population total, with proportional allocation, is

$$N \sqrt{\frac{\sum W_h \sigma_h^2}{n}} \sqrt{\left(1 - \frac{n}{N}\right)}$$

17.10—Stratified sampling for attributes. If an attribute is being sampled, the estimate appropriate to stratified sampling is

$$p_{st} = \sum W_h p_h,$$

where p_h is the sample proportion in stratum h and $W_h = N_h/N$ is the stratum weight. To find the standard error of p_{st} we substitute $p_h q_h$ for s_h^2 in the formulas previously given in section 17.8.

As an example, consider a sample of 692 families in Iowa to determine, among other things, how many had vegetable gardens in 1943. The families were arranged in three strata—urban, rural non-farm, and farm—because it was anticipated that the three groups might show differences in the frequency and size of vegetable gardens. The data are given in table 17.10.1.

The numbers of families were taken from the 1940 census. The sample was allotted roughly in proportion to the number of families per stratum, a sample of 1 per 1,000 being aimed at.

The weighted mean percentage of Iowa families having gardens was estimated as

$$\sum W_h p_h = (0.445)(72.7) + (0.230)(94.8) + (0.325)(96.6) = 85.6\%$$

TABLE 17.10.1
NUMBERS OF VEGETABLE GARDENS AMONG IOWA FAMILIES, ARRANGED IN THREE STRATA

Stratum	Number of Families N_h	Weight W_h	Number in Sample n_h	Number With Gardens	Percentage With Garden
Urban	312,393	0.445	300	218	72.7
Rural non-farm	161,077	0.230	155	147	94.8
Farm	228,354	0.325	237	229	96.6
Total	701,824	1.000	692	594	

This is practically the same as the sample mean percentage, 594/692 or 85.8%, because allocation was so close to proportional.

For the estimated variance of the estimated mean, we have

$$\Sigma W_h^2 p_h q_h / n_h = (0.445)^2 (72.7)(27.3)/300 + \text{etc.} = 1.62$$

The standard error, then, is 1.27%.

With a sample of this size, the estimated mean will be approximately normally distributed: the confidence limits may be set as

$$85.6 \pm (2)(1.27) : 83.1\% \text{ and } 88.1\%$$

For the optimum choice of the sample sizes within strata, we should take n_h proportional to $N_h \sqrt{p_h q_h / c_h}$. If the cost of sampling is about the same in all strata, as is true in many surveys, this implies that the fraction sampled, n_h / N_h , should be proportional to $\sqrt{p_h q_h}$. Now the quantity $\sqrt{p q}$ changes little as p ranges from 25% to 75%. Consequently, proportional allocation is often highly efficient in stratified sampling for attributes. The optimum allocation produces a substantial reduction in the standard error, as compared with proportional allocation, only when some of the p_h are close to zero or 100%, or when there are differential costs.

The example on vegetable gardens departs from the strict principles of stratified sampling in that the strata sizes and weights were not known exactly, being obtained from census data three years previously. Errors in the strata weights reduce the gain in precision from stratification and make the standard formulas inapplicable. It is believed that in this example these disturbances are of negligible importance. Discussions of stratification when errors in the weights are present are given in (2) and (10).

EXAMPLE 17.10.1—In stratified sampling for attributes, the optimum sample distribution, with equal costs per unit in all strata, follows from taking n_h proportional to $N_h \sqrt{p_h q_h}$. It follows that the actual value of n_h is

$$n_h = n \left\{ \frac{N_h \sqrt{p_h q_h}}{\sum N_h \sqrt{p_h q_h}} \right\}$$

In the Iowa vegetable garden survey, suppose that the p_h values found in the sample can be assumed to be the same as those in the population. Show that the optimum sample distribution gives sample sizes of 445, 115, and 132 in the respective strata, and that the standard error of the estimated percentage with gardens would then be 1.17%, as compared with 1.27% in the sample itself.

17.11—Sampling in two stages. Consider the following miscellaneous group of sampling problems: (1) a study of the vitamin A content of butter produced by creameries, (2) a study of the protein content of wheat in the wheat fields in an area, (3) a study of red blood cell counts in a population of men aged 20–30, (4) a study of insect infestation of the leaves of the trees in an orchard, and (5) a study of the number of defective teeth in third-grade children in the schools of a large city. What do these investigations have in common? First, in each study an appropriate sampling unit suggests itself naturally—the creamery, the field of wheat, the individual man, the tree, and the school. Secondly, and this is the important point, in each study the chosen sampling units can be *sub-sampled* instead of being measured completely. Indeed, sub-sampling is essential in the first three studies. No one is going to allow us to take *all* the butter produced by a creamery in order to determine vitamin A content, or *all* the wheat in a field for the protein determination, or *all* the blood in a man in order to make a complete count of his red cells. In the insect infestation study, it might be feasible, although tedious, to examine *all* leaves on any selected tree. If the insect distribution is spotty, however, we would probably decide to take only a small sample of leaves from any selected tree in order to include more trees. In the dental study—we could take all the third-grade children in any selected school or we could cover a larger sample of schools by examining only a sample of children from the third grade in each selected school.

This type of sampling is called *sampling in two stages*, or sometimes *sub-sampling*. The first stage is the selection of a sample of *primary sampling units*—the creameries, wheat fields, and so on. The second stage is the taking of a *sub-sample* of *second-stage units*, or *sub-units*, from each selected primary unit.

As illustrated by these examples, the two-stage method is sometimes the only practicable way in which the sampling can be done. Even when there is a choice between sub-sampling the units and measuring them completely, two-stage sampling gives the sampler greater scope, since he can choose both the size of the sample of primary units and the size of the sample that is taken from a primary unit. In some applications an important advantage of two-stage sampling is that it facilitates the problem of listing the population. Often it is relatively easy to obtain a list of the primary units, but difficult or expensive to list all the sub-units. To list the trees in an orchard and draw a sample of them is usually simple, but the problem of making a random selection of the leaves on a tree may be very troublesome. With two-stage sampling this problem is faced only for those trees that are in the sample. No complete listing of all leaves in the orchard is required.

In the discussion of two-stage sampling we assume at first that the primary units are of approximately the same size. A simple random sample of n_1 primary units is drawn, and the same number n_2 of sub-units is selected from each primary unit in the sample. The estimated standard error of the sample mean \bar{y} per sub-unit is then given by the formula

$$s_{\bar{y}} = \frac{1}{\sqrt{n_1}} \sqrt{\frac{\sum (\bar{y}_i - \bar{y})^2}{n_1 - 1}},$$

where \bar{y}_i is the mean per sub-unit in the i th primary unit. This formula does not include the finite population correction, but is reliable enough provided that the sample contains less than 10% of all primary units. Note that the formula makes no use of the individual observations on the sub-units, but only of the primary unit means \bar{y}_i . If the sub-samples are taken for a chemical analysis, a common practice is to composite the sub-sample and make one chemical determination for each primary unit. With data of this kind we can still calculate $s_{\bar{y}}$.

In section 10.13 you learned about the "components of variance" technique, and applied it to a problem in two-stage sampling. The data were concentrations of calcium in turnip greens, four determinations being made for each of three leaves. The leaf can be regarded as the primary sampling unit, and the individual determination as the sub-unit. By applying the components of variance technique, you were able to see how the variance of the sample mean was affected by variation between determinations on the same leaf and by variation from leaf to leaf. You could also predict how the variance of the sample mean would change with different numbers of leaves and of determinations per leaf in the experiment.

Since this technique is of wide utility in two-stage sampling, we shall repeat some of the results. The observation on any sub-unit is considered to be the sum of two independent terms. One term, associated with the primary unit, has the same value for all second-stage units in the primary unit, and varies from one primary unit to another with variance σ_1^2 . The second term, which serves to measure differences between second-stage units, varies independently from one sub-unit to another with variance σ_2^2 . Suppose that a sample consists of n_1 primary units, from each of which n_2 sub-units are drawn. Then the sample as a whole contains n_1 independent values of the first term, whereas it contains $n_1 n_2$ independent values of the second term. Hence the variance of the sample mean \bar{y} per sub-unit is

$$\sigma_{\bar{y}}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1 n_2}$$

The two components of variance, σ_1^2 and σ_2^2 , can be estimated from the analysis of variance of a two-stage sample that has been taken. Table 17.11:1 gives the analysis of variance for a study by Immer (6), whose

object was to develop a sampling technique for the determination of the sugar percentage in field experiments on sugar beets. Ten beets were chosen from each of 100 plots in a uniformity trial, the plots being the primary units. The sugar percentage was obtained separately for each beet. In order to simulate conditions in field experiments, the Between plots mean square was computed as the mean square between plots within blocks of 5 plots. This mean square gives the experimental error variance that would apply in a randomized blocks experiments with 5 treatments.

TABLE 17.11.1
ANALYSIS OF VARIANCE OF SUGAR PERCENTAGE OF BEETS (ON A SINGLE-BEET BASIS)

Source of Variation	Degrees of Freedom	Mean Square	Parameters Estimated
Between plots (primary units)	80	2.9254	$\sigma_2^2 + 10 \sigma_1^2$
Between beets (sub-units) within plots	900	2.1374	σ_2^2

The estimate of σ_1^2 , the Between plots component of variance, is

$$s_1^2 = \frac{2.9254 - 2.1374}{10} = 0.0788,$$

the divisor 10 being the number of beets (sub-units) taken per plot. As an estimate of σ_2^2 , the within-plots component, we have

$$s_2^2 = 2.1374$$

Hence, if a new experiment is to consist of n_1 replications, with r beets sampled from each plot, the predicted variance of a treatment mean is

$$s_{\bar{y}}^2 = \frac{0.0788}{n_1} + \frac{2.1374}{n_1 n_2}$$

We shall illustrate two of the questions that can be answered from these data. How accurate are the treatment means in an experiment with 6 replications and 5 beets per plot? For this experiment we would expect

$$s_{\bar{y}} = \sqrt{\left(\frac{0.0788}{6} + \frac{2.1374}{30} \right)} = 0.29\%$$

The sugar percentage figure for a treatment mean would be correct to within $\pm(2)(0.29)$ or 0.58%, with 95% confidence, assuming \bar{y} approximately normally distributed.

If the standard error of a treatment mean is not to exceed 0.2%, what combinations of n_1 and n_2 are allowable? We must have

$$\frac{0.0788}{n_1} + \frac{2.1374}{n_1 n_2} = (0.2)^2 = 0.04$$

Since n_1 and n_2 are whole numbers, they will not satisfy this equation exactly: we must make sure that the left side of the equation does not exceed 0.04. You can verify that with 4 replications ($n_1 = 4$), there must be 27 beets per plot; with 8 replications, 9 beets per plot are sufficient; and with 10 replications, 7 beets per plot. As one would expect, the intensity of sub-sampling decreases as the intensity of sampling is increased. The total size of sample also decreases from 108 beets when $n_1 = 4$ to 70 beets when $n_1 = 10$.

17.12—The allocation of resources in two-stage sampling. The last example illustrates a general property of two-stage samples. The same standard error can be attained for the sample mean by using various combinations of values of n_1 and n_2 . Which of these choices is the best? The answer depends, naturally, on the cost of adding an extra primary unit to the sample (in this case an extra replication) relative to that of adding an extra sub-unit in each primary unit (in this case an extra beet in each plot). Similarly, in the turnip greens example (section 10.13, page 280) the best sampling plan depends on the relative costs of taking an extra leaf and of making an extra determination per leaf. Obviously, if it is cheap to add primary units to the sample but expensive to add sub-units, the most economical plan will be to have many primary units and few (perhaps only one) sub-units per primary unit. For a general solution to this problem, however, we require a more exact formulation of the costs of various alternative plans.

In many sub-sampling studies the cost of the sample (apart from fixed overhead costs) can be approximated by a relation of the form

$$\text{cost} = c_1 n_1 + c_2 n_1 n_2$$

The factor c_1 is the average cost per primary unit of those elements of cost that depend solely on the number of primary units and not on the amount of sub-sampling. The factor c_2 , on the other hand, is the average cost per sub-unit of those constituents of cost that are directly proportional to the total number of sub-units.

If advance estimates of these constituents of cost are made from a preliminary study, an efficient job of selecting the best amounts of sampling and sub-sampling can be done. The problem may be posed in two different ways. In some studies we specify the desired variance V for the sample mean, and would like to attain this as cheaply as possible. In other applications the total cost C that must not be exceeded is imposed upon us, and we want to get as small a value of V as we can for this outlay. These two problems have basically the same solution. In either case we want to minimize the product

$$VC = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1 n_2} \right) (c_1 n_1 + c_2 n_1 n_2)$$

Upon expansion, this becomes

$$VC = (s_1^2 c_1 + s_2^2 c_2) + n_2 s_1^2 c_2 + \frac{s_2^2 c_1}{n_2}$$

It can be shown that this expression has its smallest value when

$$n_2 = \sqrt{\frac{c_1 s_2^2}{c_2 s_1^2}}$$

This result gives an estimate of the best number of sub-units (beets) per primary unit (plot). The value of n_1 is found by solving either the cost equation or the variance equation for n_1 , depending on whether cost or variance has been preassigned.

In the sugar beet example we had $s_1^2 = 0.0788$, $s_2^2 = 2.1374$, from which

$$n_2 = \sqrt{\frac{2.1374}{0.0788}} \sqrt{\frac{c_1}{c_2}} = 5.2 \sqrt{\frac{c_1}{c_2}}$$

In this study, cost data were not reported. If c_1 were to include the cost of the land and the field operations required to produce one plot, it would be much greater than c_2 . Evidently a fairly large number of beets per plot would be advisable. In practice, factors other than the sugar percentage determinations must also be taken into account in deciding on costs and number of replications in sugar beet experiments.

In the turnip greens example (section 10.13, page 280), n_1 is the number of leaves and n_2 the number of determinations of calcium concentration per leaf. Also, in the present notation,

$$\begin{aligned} s_1^2 &= s_A^2 = 0.0724 \\ s_2^2 &= s^2 = 0.0066 \end{aligned}$$

Hence, the most economical number of determinations per leaf is estimated to be

$$n_2 = \sqrt{\frac{c_1 s_2^2}{c_2 s_1^2}} = \sqrt{\frac{0.0066}{0.0724}} \sqrt{\frac{c_1}{c_2}} = 0.30 \sqrt{\frac{c_1}{c_2}}$$

In practice, n_2 must be a whole number, and the smallest value it can have is 1. This equation shows that $n_2 = 1$, i.e., one determination per leaf, unless c_1 is at least 25 times c_2 . Actually, since c_2 includes the cost of the chemical determinations, it is likely to be greater than c_1 . The relatively large variation among leaves and the cost considerations both point to the choice of one determination per leaf.

This example also illustrates that a choice of n_2 can often be made from the equation even when information about relative costs is not too definite. This is because the equation often leads to the same value of n_2 for a wide range of ratios of c_1 to c_2 . Brooks (14) gives helpful tables for

this situation. The values of n_2 are subject to sampling errors; for a discussion, see (2).

In section 10.14 you studied an example of *three-stage sampling* of turnip green plants. The first stage was represented by plants, the second by leaves within plants, and the third by determinations within a leaf. In the notation of this section, the estimated variance of the sample mean is

$$s_{\bar{y}}^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_1 n_2} + \frac{s_3^2}{n_1 n_2 n_3}$$

Copying the equation given in section 10.14, we have

$$s_{\bar{y}}^2 = \frac{0.3652}{n_1} + \frac{0.1610}{n_1 n_2} + \frac{0.0067}{n_1 n_2 n_3}$$

To find the most economical values of n_1 , n_2 , and n_3 , we set up a cost equation of the form

$$\text{cost} = c_1 n_1 + c_2 n_1 n_2 + c_3 n_1 n_2 n_3$$

and proceed to minimize the product of the variance and the cost as before. The solutions are

$$n_2 = \sqrt{\frac{c_1 s_2^2}{c_2 s_1^2}}, \quad n_3 = \sqrt{\frac{c_2 s_3^2}{c_3 s_2^2}},$$

while n_1 is found by solving either the cost or the variance equation. Note that the formula for n_2 is the same in three-stage as in two-stage sampling, and that the formula for n_3 is the natural extension of that for n_2 . Putting in the numerical values of the variance components, we obtain

$$n_2 = \sqrt{\frac{c_1(0.1610)}{c_2(0.3652)}} = 0.66 \sqrt{\frac{c_1}{c_2}}, \quad n_3 = \sqrt{\frac{c_2(0.0067)}{c_3(0.1610)}} = 0.20 \sqrt{\frac{c_2}{c_3}}$$

Since the computed value of n_3 would be less than 1 for any likely value of c_2/c_3 , more than one determination per leaf is uneconomical. The optimum number n_2 of leaves per plant depends on the ratio c_1/c_2 . This will vary with the conditions of experimentation. If many plants are being grown for some other purpose, so that ample numbers are available for sampling, c_1 includes only the extra costs involved in collecting a sample from many plants instead of a few plants. In this event the optimum n_2 might also turn out to be 1. If the cost of growing extra plants is to be included in c_1 , the optimum n_2 might be higher than 1.

EXAMPLE 17 12.1—This is the analysis of variance, on a single sub-sample basis, for wheat yield and percentage of protein from data collected in a wheat sampling survey in Kansas in 1939 (25)

Source of Variation	Yield (Bushels Per Acre)		Protein (%)	
	Degrees of Freedom	Mean Square	Degrees of Freedom	Mean Square
Fields	659	434.52	659	21.388
Samples within fields	660	67.54	609	2.870

Two sub-samples were taken at random from each of 660 fields. Calculate the components of variance for yield. Ans. $s_1^2 = 183.49$, $s_2^2 = 67.54$. Note: Some of the protein figures were evidently not recorded separately for each sub-sample, since there are only 609 d.f. within fields.

EXAMPLE 17.12.2--For yield, estimate the variance of the sample mean for samples consisting of (i) 1 sub-sample from each of 800 fields, (ii) 2 sub-samples from each of 400 fields, (iii) 8 samples from each of 100 fields. Ans. (i) 0.313, (ii) 0.543, (iii) 1.919.

EXAMPLE 17.12.3—With 2 sub-samples per field, it is desired to take enough fields so that the standard error of the mean yield will be not more than $1/2$ bushel, and at the same time the standard error of the mean protein percentage will be not more than $\frac{1}{8}\%$. How many fields are required? Ans. about 870.

EXAMPLE 17.12.4—Suppose that it takes on the average 1 man-hour to locate and pace a field that is to be sampled. A single protein determination is to be made on the bulked sub-samples from any field. The cost of a determination is equivalent to 1 man-hour. It takes 15 minutes to locate, cut, and tie a sub-sample. From these data and the analysis of variance for protein percentage (example 17.12.1), compute the variance-cost product, VC , for each value of n_2 from 1 to 5. What is the most economical number of sub-samples per field? Ans. 2. How much more does it cost, for the same V , if 4 sub-samples per field are used? Ans. 12%.

17.13—Selection with probability proportional to size. In many important sampling problems, the natural primary sampling units vary in size. In national surveys conducted to obtain information about the characteristics of the population, the primary unit is often an administrative area (e.g., similar to a county). A relatively large unit of this type cuts down travel costs and makes supervision and control of the field work more manageable. Such units often vary substantially in the number of people they contain. A sample of the houses in a town may use blocks as first-stage units, the number of houses per block ranging from 0 to 40. Similarly, schools, hospitals, and factories all contain different numbers of individuals.

With primary units like this, the between-primary-unit variances of the principal measurements may be large; for example, some counties are relatively wealthy and some are poor. In these circumstances, Hansen and Hurwitz (15) pointed out the advantages of selecting primary units with probabilities proportional to their sizes. To illustrate, consider a population of three schools, having 600, 300, and 100 children. The objective is to estimate the population mean per child for some characteristic. The means per child in the three schools are $\bar{Y}_1 = 2$, $\bar{Y}_2 = 4$, $\bar{Y}_3 = 1$. Hence, the population mean per child is

$$\bar{Y} = [(600)(2) + (300)(4) + (100)(1)]/1,000 = 2.5$$

To simplify things further, suppose that only one school is to be chosen, and that the variation in Y between children within the same school is negligible. It follows that we need not specify how the second-stage sample of children from a school is to be drawn, since any sample gives the correct mean for the chosen school.

In selecting the school with probability proportional to size, (*pps*), the three schools receive probabilities 0.6, 0.3, and 0.1, respectively, of being drawn. We shall compare the mean square error of the estimate given by this method with that given by selecting the schools with equal probabilities. Table 17.13.1 contains the calculations.

TABLE 17.13.1
SELECTION OF A SCHOOL WITH PROBABILITY PROPORTIONAL TO SIZE

School	No. of Children	Probability of Selection π_i	Mean per Child \bar{Y}_i	Error of Estimate $\bar{Y}_i - \bar{Y}$	$(\bar{Y}_i - \bar{Y})^2$
1	600	0.6	2	-0.5	0.25
2	300	0.3	4	+1.5	2.25
3	100	0.1	1	-1.5	2.25
Population	1,000	1.0	2.5		

If the first school is selected, its estimate is in error by $(2.0 - 2.5) = -0.5$, and so on. These errors and their squares appear in the two right-hand columns of table 17.13.1. In repeated sampling with probability proportional to size, the first school is drawn 60% of the time, the second school 30%, and the third school 10%. The mean square error is therefore

$$M.S.E_{pps} = (0.6)(0.25) + (0.3)(2.25) + (0.1)(2.25) = 1.05$$

If, alternatively, the schools are drawn with equal probability, the $M.S.E$ is

$$M.S.E_{eq} = \frac{1}{3}[(0.25) + (2.25) + (2.25)] = 1.58$$

This $M.S.E$ is about 50% higher than that given by *pps* selection.

You may ask: Does this result depend on the choice or the order of the means, 2, 4, 1, assigned to schools 1, 2, and 3? The answer is yes. With means 4, 2, 1, you will find $M.S.E_{pps} = 1.29$, $M.S.E_{eq} = 2.14$, the latter being 66% higher. Over the six possible orders of the numbers 1, 2, 4, the ratio $M.S.E_{eq}/M.S.E_{pps}$ varies from 0.93 to 2.52. However, the ratio of the averages $\overline{M.S.E}_{eq}/\overline{M.S.E}_{pps}$, taken over all six possible orders, does not depend on the numbers 1, 2, 4. With N primary units in the population, this ratio is

$$\frac{\overline{M.S.E}_{eq}}{\overline{M.S.E}_{pps}} = \frac{(N-1) + N \sum (\pi_i - \bar{\pi})^2}{(N-1) - N \sum (\pi_i - \bar{\pi})^2}$$

where π_i is the probability of selection (relative size) of the i th school. Clearly, this ratio exceeds one unless all π_i are equal; that is, all schools are the same size.

The reason why it usually pays to select large units with higher probabilities is that the population mean depends more on the means of the large units than on those of the small units. The large units are therefore likely to give better estimates.

With two-stage sampling, a simple method is to select n primary units with *pps* and take an *equal number* of sub-units (e.g., children) in every selected primary unit. This method gives every sub-unit in the population the same chance of being in the sample. The sample mean per sub-unit \bar{y} is an unbiased estimate of the corresponding population mean, and its estimated variance is obtained by the simple formula

$$s_{\bar{y}}^2 = \sum (\bar{y}_i - \bar{\bar{y}})^2 / n(n-1), \quad (17.13.1)$$

where \bar{y}_i is the mean of the sample from the i th primary unit.

We have illustrated only the simplest case. Formula 17.13.1 assumes that the n units are selected with replacement (i.e., that a unit can be chosen more than once). Some complications arise when we select units without replacement. Often, the sizes of the units are not known exactly, and have to be estimated in advance. Considerations of cost or of the structure of variability in the population may lead to the selection of units with probabilities that are unequal, but are proportional to some quantity other than the sizes. For details, see the references. In extensive surveys, multistage sampling with unequal probabilities of selection of primary units is the commonest method in current practice.

17.14—Ratio and regression estimates. The *ratio estimate* is a different way of estimating population totals (or means) that is useful in many sampling problems. Suppose that you have taken a sample in order to estimate the population total of a variable, Y , and that a complete count of the population was made on some previous occasion. Let X denote the value of the variable on the previous occasion. You might then compute the ratio

$$R = \frac{\sum Y}{\sum X},$$

where the sums are taken over the sample. This ratio is an estimate of the present level of the variate relative to that on the previous occasion. On multiplying the ratio by the known population total on the previous

occasion (i.e., by the population total of X), you obtain the ratio estimate of the population total of Y . Clearly, if the relative change is about the same on all sampling units, the ratio R will be accurate and the estimate of the population total will be a good one.

The ratio estimate can also be used when X is some other kind of supplementary variable. The conditions for a successful application of this estimate are that the ratio Y/X should be relatively constant over the population and that the population total of X should be known. Consider an estimate of the total amount of a crop, just after harvest, made from a sample of farms in some region. For each farm in the sample we record the total yield, Y , and the total acreage, X , of that crop. In this case the ratio, $R = \Sigma Y / \Sigma X$, is the sample estimate of the mean yield per acre. This is multiplied by the total acreage of the crop in the region, which would have to be known accurately from some other source. This estimate will be precise if the mean yield per acre varies little from farm to farm.

The estimated standard error of the ratio estimate \hat{Y}_R of the population total from a simple random sample of size n is, approximately,

$$s(\hat{Y}_R) = N \sqrt{\frac{\Sigma(Y - RX)^2}{n(n-1)}}$$

The ratio estimate is not always more precise than the simpler estimate $N\bar{y}$ (number of units in population \times sample mean). It has been shown that the ratio estimate is more precise only if ρ , the correlation coefficient between Y and X , exceeds $C_X/2C_Y$, where the C 's are the coefficients of variation. Consequently, ratio estimates must not be used indiscriminately, although in appropriate circumstances they produce large gains in precision.

Sometimes the purpose of the sampling is to estimate a ratio, e.g. ratio of dry weight to total weight or ratio of clean wool to total wool. The estimated standard error of the estimate is then

$$s(R) = \frac{1}{\bar{x}} \sqrt{\frac{\Sigma(Y - RX)^2}{n(n-1)}}$$

This formula has already been given (in a different notation) at the end of section 17.5, where the estimation of proportions from cluster sampling was discussed.

In chapter 6 the linear regression of Y on X and its sample estimate,

$$\hat{Y} = \bar{y} + bx,$$

were discussed. With an auxiliary variable, X , you may find that when you plot Y against X from the sample data, the points appear to lie close to a straight line, but the line does not go through the origin. This implies that the ratio Y/X is not constant over the sample. As pointed out in section 6.19, it is then advisable to use a linear regression estimate instead

of the ratio estimate. For the population total of Y , the linear regression estimate is

$$N\hat{Y} = N\{\bar{y} + b(\bar{X} - \bar{x})\},$$

where \bar{X} is the population mean of X . The term inside the brackets is the sample mean, \bar{y} , adjusted for regression. To see this, suppose that you have taken a sample in which $\bar{y} = 2.35$, $\bar{x} = 1.70$, $\bar{X} = 1.92$, $b = +0.4$. Your first estimate of the population mean would be $\bar{y} = 2.35$. But in the sample the mean value of X is too low by an amount $(1.92 - 1.70) = 0.22$. Further, the value of b tells you that unit increase in X is accompanied, on the average, by $+0.4$ unit increase in Y . Hence, to correct for the low value of the mean of X , you increase the sample mean by the amount $(+0.4)(0.22)$. Thus the adjusted value of \bar{y} is

$$2.35 + (+0.4)(0.22) = 2.44 = \bar{y} + b(\bar{X} - \bar{x})$$

To estimate the population total, this value is multiplied by N , the number of sampling units in the population.

The standard error of the estimated population total is, approximately,

$$s_{N\bar{y}} = Ns_{y \cdot x} \sqrt{\left(\frac{1}{n} + \frac{(\bar{X} - \bar{x})^2}{\Sigma x^2}\right)}$$

If a finite population correction is required in the standard error formulas presented in this section, insert the factor $\sqrt{(1 - \phi)}$. In finite populations the ratio and regression estimates are both slightly biased, but the bias is seldom important in practice.

17.15—Further reading. The general books on sample surveys that have become standard, (2), (3), (4), (5), (13), involve roughly the same level of mathematical difficulty and knowledge of statistics. Reference (3) is oriented towards applications in business, and reference (13) towards those in agriculture. Another good book for agricultural applications, at a lower mathematical level, is (16).

Useful short books are (17), an informal, popular account of some of the interesting applications of survey methods, (18), which conducts the reader painlessly through the principal results in probability sampling at about the mathematical level of this chapter, and (19), which discusses the technique of constructing interview questions.

Books and papers have also begun to appear on some of the common specific types of application. For sampling a town under U.S. conditions, with the block as primary sampling unit, references (20) and (21) are recommended. Reference (22), intended primarily for surveys by health agencies to check on the immunization status of children, gives instructions for the sampling of attributes in local areas, while (24) deals with the sampling of hospitals and patients. Much helpful advice on the use of

sampling in agricultural censuses is found in (23). Sampling techniques for estimating the volume of timber of the principal types and age-classes in forestry are summarized in (11), while (9) reviews the difficult problem of estimating wildlife populations.

REFERENCES

1. A. R. CLAPHAM. *J. Agric. Sci.*, 19:214 (1929).
2. W. G. COCHRAN. *Sampling Techniques*. Wiley, 2nd ed. New York (1963).
3. W. EDWARDS DEMING. *Sample Design in Business Research*. Wiley, New York (1960).
4. M. H. HANSEN, W. N. HURWITZ, and W. G. MADOW. *Sample Survey Methods and Theory*. Wiley, New York (1953).
5. L. KISH. *Survey Sampling*. Wiley, New York (1965).
6. F. R. IMMER. *J. Agric. Res.*, 44:633 (1932).
7. A. J. KING, D. E. McCARTY, and M. McPEAK. USDA Tech. Bull. 814 (1942).
8. J. A. RIGNEY and J. FIELDING REED. *J. Amer. Soc. Agron.*, 39:26 (1947).
9. L. W. SCATTERGOOD. Chap. 20 in *Statistics and Mathematics in Biology*. Iowa State College Press (1954).
10. F. F. STEPHAN. *J. Marketing*, 6:38 (1941).
11. A. A. HASEL. Chap. 19 in *Statistics and Mathematics in Biology*. Iowa State College Press (1954).
12. Q. M. WEST. Mimeographed Report, Cornell Univ. Agric. Exp. Sta. (1951).
13. F. YATES. *Sampling Methods for Censuses and Surveys*, 3rd ed. Charles Griffin, London (1960).
14. S. BROOKS. *J. Amer. Statist. Ass.*, 50:398 (1955).
15. M. H. HANSEN and W. N. HURWITZ. *Ann. Math. Statist.*, 14:333 (1943).
16. M. R. SAMPFORD. *An Introduction to Sampling Theory*. Oliver and Boyd, Edinburgh (1962).
17. M. J. SLONIM. *Sampling in a Nutshell*. Simon and Schuster, New York (1960).
18. A. STUART. *Basic Ideas of Scientific Sampling*. Charles Griffin, London (1962).
19. S. L. PAYNE. *The Art of Asking Questions*. Princeton University Press (1951).
20. T. D. WOOLSEY. "Sampling Methods for a Small Household Survey." *Public Health Monographs*, No. 40 (1956).
21. L. KISH. *Amer. Soc. Rev.*, 17:761 (1952).
22. R. E. SERFLING and I. L. SHERMAN. *Attribute Sampling Methods*. U.S. Govt. Printing Office, Washington, D.C. (1965).
23. S. S. ZARCOVICH. *Sampling Methods and Census*. FAO, Rome (1965).
24. I. HESS, D. C. RIEDEL, and T. B. FITZPATRICK. *Probability Sampling of Hospitals and Patients*. University of Michigan, Ann Arbor, Mich. (1961).
25. A. J. KING and D. E. McCARTY. *J. Marketing*, 6:462 (1941).

List of Appendix Tables.

A 1	Random digits	543
A 2	Normal distribution, ordinates	547
A 3	Normal distribution, cumulative frequency	548
A 4	Student's t , percentage points	549
A 5	Chi-square, χ^2 , percentage points	550
A 6 (i)	Test for skewness, 5% and 1% points of g_1	552
A 6 (ii)	Test for kurtosis, 5% and 1% points of g_2	552
A 7 (i)	t_r , range analog of t , 10%, 5%, 2%, and 1% points	553
A 7 (ii)	Two-sample range analog of t , 10%, 5%, 2%, and 1% points	554
A 8	Sign test, 10%, 5%, and 1% points	554
A 9	Signed rank test, 5% and 1% points	555
A 10	Two-sample signed rank test, 5% and 1% points	555
A 11	Correlation coefficient, r , 5% and 1% points	557
A 12	Transformed correlations, z in terms of r	558
A 13	Transformed correlations, r in terms of z	559
A 14 (i)	F , variance ratio, 5% and 1% points	560
A 14 (ii)	F , variance ratio, 25%, 10%, 2.5%, and 0.5% points	564
A 15	Studentized range, Q , 5% points	568
A 16	Angular transformation. Angle = $\arcsin \sqrt{\text{percentage}}$	569
A 17	Orthogonal polynomial values	572
A 18	Square roots	573

Notes

Interpolation. In analyses of data and in working the examples in this book, use of the nearest entry in any Appendix table is accurate enough in most cases. The following examples illustrate linear interpolation, which will sometimes be needed.

1. Find the 5% significance level of χ^2 for 34 degrees of freedom. For $P = 0.050$, table A 5 gives

$d.f.$	30	34	40
χ^2	43.77	?	55.76

Calculate $(34 - 30)/(40 - 30) = 0.4$. Since

$$34 = 30 + 0.4(40 - 30)$$

the required value of χ^2 is

$$43.77 + 0.4(55.76 - 43.77) = 43.77 + 0.4(11.99) = 48.57$$

Alternatively, this value can be computed as

$$(0.4)\chi_{40}^2 + (0.6)\chi_{30}^2 = (0.4)(55.76) + (0.6)(43.77) = 48.57$$

542 **Appendix Tables**

Note that 0.4 multiplies χ^2_{40} , not χ^2_{30} .

2. An analysis gave an F value of 2.04 for 3 and 18 $d.f.$ Find the significance probability. For 3 and 18 $d.f.$, table A 14, part II, gives the following entries:

P	0.25	?	0.10
F	1.49	2.04	2.42

Calculate $(2.04 - 1.49)/(2.42 - 1.49) = 0.55/0.93 = 0.59$. By the alternative method in the preceding example.

$$P = (0.59)(0.10) + (0.41)(0.25) = 0.16$$

Finding Square Roots. Table A 18 is a table of square roots. To save space the entries jump by 0.02 instead of 0.01, but interpolation will rarely be necessary. With very large or very small numbers, mistakes in finding square roots are common. The following examples should clarify the procedure.

Step	(1)	(2)	(3)	(4)
Number	Mark Off	Column Read	Reading	Square Root
6,028.0	60,28.0	$\sqrt{10n}$	7.76	77.6
397 2	3,97 2	\sqrt{n}	1.99	19.9
46.38	46.38	$\sqrt{10n}$	6.81	6.81
0.194	0.19,4	$\sqrt{10n}$	4.40	0.440
0 000893	0.00,08,93	\sqrt{n}	2.99	0.0299

In step (1), mark off the digits *in twos* to the right or left of the decimal point. Step (2) tells which column of the square root table is to be read. With 3,97.2 and 0.00,08,93 read the \sqrt{n} column, because there is a *single* digit (3 or 8) to the left of the first comma that has any non-zero digits to its left. If there are *two* digits to the left of the first comma, as in 60,28.0, read the $\sqrt{10n}$ column. Step (3) gives the reading, taken directly from the nearest entry in the table.

The final step (4) finds the actual square roots. Consider, first, numbers greater than 1. If column (1) has no comma to the left of the decimal, as with 46.38, the square root has one digit to the left of the decimal. If column (1) has one comma to the left of the decimal, as with 60,28.0 and 3,97.2 the square root has two digits to the left of the decimal, and so on. With numbers smaller than 1, replace any pair 00 to the right of the decimal by a single 0. Thus, the square root of 0.00,08,93 is 0.0299 as shown. The square root of 0.00,00,08,93 is 0.00299.

TABLE A 1
TEN THOUSAND RANDOMLY ASSORTED DIGITS

	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927
02	85941	40756	82414	02015	13858	78030	16269	65978	01385	15345
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976
06	28357	94070	20652	35774	16249	75019	21145	05217	47286	76305
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491
15	98614	75993	84460	62846	59844	14922	48730	73443	48167	34770
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771
18	90801	21472	42815	77408	37390	76766	52615	32141	30268	18106
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067
20	75884	12952	84318	95108	72305	64620	91318	89872	45375	85436
21	16777	37116	58550	42958	21460	43910	01175	87894	81378	10620
22	46230	43877	80207	88877	89380	32992	91380	03164	98656	59337
23	42902	66892	46134	01432	94710	23474	20423	60137	60609	13119
24	81007	00333	39693	28039	10154	95425	39220	19774	31782	49037
25	68089	01122	51111	72373	06902	74373	96199	97017	41273	21546
26	20411	67081	89950	16944	93054	87687	96693	87236	77054	33848
27	58212	13160	06468	15718	82627	76999	05999	58680	96739	63700
28	70577	42866	24969	61210	76046	67699	42054	12696	93758	03283
29	94522	74358	71659	62038	79643	79169	44741	05437	39038	13163
30	42626	86819	85651	88678	17401	03252	99547	32404	17918	62880
31	16051	33763	57194	16752	54450	19031	58580	47629	54132	60631
32	08244	27647	33851	44705	94211	46716	11738	55784	95374	72655
33	59497	04392	09419	89964	51211	04894	72882	17805	21896	83864
34	97155	13428	40293	09985	58434	01412	69124	82171	59058	82859
35	98409	66162	95763	47420	20792	61527	20441	39435	11859	41567
36	45476	84882	65109	96597	25930	66790	65706	61203	53634	22557
37	89300	69700	50741	30329	11658	23166	05400	66669	48708	03887
38	50051	95137	91631	66315	91428	12275	24816	68091	71710	33258
39	31753	85178	31310	89642	98364	02306	24617	09609	83942	22716
40	79152	53829	77250	20190	56535	18760	69942	77448	33278	48805
41	44560	38750	83635	56540	64900	42912	13953	79149	18710	68618
42	68328	83378	63369	71381	39564	05615	42451	64559	97501	65747
43	46939	38689	58625	08342	30459	85863	20781	09284	26333	91777
44	83544	86141	15707	96256	23068	13782	08467	89469	93842	55349
45	91621	00881	04900	54224	46177	55309	17852	27491	89415	23466
46	91896	67126	04151	03795	59077	11848	12630	98375	52068	60142
47	55751	62515	21108	80830	02263	29303	37204	96926	30506	09808
48	85156	87689	95493	88842	00664	55017	55539	17771	69448	87530
49	07521	56898	12236	60277	39102	62315	12239	07105	11844	01117

TABLE A 1—(Continued)

	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
00	59391	58030	52098	82718	87024	82848	04190	96574	90464	29065
01	99567	76364	77204	04615	27062	96621	43918	01896	83991	51141
02	10363	97518	51400	25670	98342	61891	27101	37855	06235	33316
03	86859	19558	64432	16706	99612	59798	32803	67708	15297	28612
04	11258	24591	36863	55368	31721	94335	34936	02566	80972	08188
05	95068	88628	35911	14530	33020	80428	39936	31855	34334	64865
06	54463	47237	73800	91017	36239	71824	83671	39892	60518	37092
07	16874	62677	57412	13215	31389	62233	80827	73917	82802	84420
08	92494	63157	76593	91316	03505	72389	96363	52887	01087	66091
09	15669	56689	35682	40844	53256	81872	35213	09840	34471	74441
10	99116	75486	84989	23476	52967	67104	39495	39100	17217	74073
11	15696	10703	65178	90637	63110	17622	53988	71087	84148	11670
12	97720	15369	51269	69620	03388	13699	33423	67453	43269	56720
13	11666	13841	71681	98000	35979	39719	81899	07449	47985	46967
14	71628	73130	78783	75691	41632	09847	61547	18707	85489	69944
15	40501	51089	99943	91843	41995	88931	73631	69361	05375	15417
16	22518	55576	98215	82068	10798	86211	36584	67466	69373	40054
17	75112	30485	62173	02132	14878	92879	22281	16783	86352	00077
18	80327	02671	98191	84342	90813	49268	95441	15496	20168	09271
19	60251	45548	02146	05597	48228	81366	34598	72856	66762	17002
20	57430	82270	10421	00540	43648	75888	66049	21511	47676	33444
21	73528	39559	34434	88596	54086	71693	43132	14414	79949	85193
22	25991	65959	70769	64721	86413	33475	42740	06175	82758	66248
23	78388	16638	09134	59980	63806	48472	39318	35434	24057	74739
24	12477	09965	96657	57994	59439	76330	24596	77515	09577	91871
25	83266	32883	42451	15579	38155	29793	40914	65990	16255	17777
26	76970	80876	10237	39515	79152	74798	39357	09054	73579	92359
27	37074	65198	44785	68624	98336	84481	97610	78735	46703	98265
28	83712	06514	30101	78295	54656	85417	43189	60048	72781	72606
29	20287	56862	69727	94443	64936	08366	27227	05158	50326	59566
30	74261	32592	86538	27041	65172	85532	07571	80609	39285	65340
31	64081	49863	08478	96001	18888	14810	70545	89755	59064	07210
32	05617	75818	47750	67814	29575	10526	66192	44464	27058	40467
33	26793	74951	95466	74307	13330	42664	85515	20632	05497	33625
34	65988	72850	48737	54719	52056	01596	03845	35067	03134	70322
35	27366	42271	44300	73399	21105	03280	73457	43093	05192	48657
36	56760	10909	98147	34736	33863	95256	12731	66598	50771	83665
37	72880	43338	93643	58904	59543	23943	11231	83268	65938	81581
38	77888	38100	03062	58103	47961	83841	25878	23746	55903	44115
39	28440	07819	21580	51459	47971	29882	13990	29226	23608	15873
40	63525	94441	77033	12147	51054	49955	58312	76923	96071	05813
41	47606	93410	16359	89033	89696	47231	64498	31776	05383	39902
42	52669	45030	96279	14709	52372	87832	02735	50803	72744	88208
43	16738	60159	07425	62369	07515	82721	37875	71153	21315	00132
44	59348	11695	45751	15865	74739	05572	32688	20271	65128	14551
45	12900	71775	29845	60774	94924	21810	38636	33717	67598	82521
46	75086	23537	49939	33595	13484	97588	28617	17979	70749	35234
47	99495	51434	29181	09993	38190	42553	68922	52125	91077	40197
48	26075	31671	45386	36583	93459	48599	52022	41330	60651	91321
49	13636	93596	23377	51133	95126	61496	42474	45141	46660	42338

TABLE A 1—(Continued)

	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
50	64249	63664	39652	40646	97306	31741	07294	84149	46797	82487
51	26538	44249	04050	48174	65570	44072	40192	51153	11397	58212
52	05845	00512	78630	55328	18116	69296	91705	86224	29503	57071
53	74897	68373	67359	51014	33510	83048	17056	72506	82949	54600
54	20872	54570	35017	88132	25730	22626	86723	91691	13191	77212
55	31432	96156	89177	75541	81355	24480	77243	76690	42507	84362
56	66890	61505	01240	00660	05873	13568	76082	79172	57913	93448
57	41894	57790	79970	33106	86904	48119	52503	24130	72824	21627
58	11303	87118	81471	52936	08555	28420	49416	44448	04269	27029
59	54374	57325	16947	45356	78371	10563	97191	53798	12693	27928
60	64852	34421	61046	90849	13966	39810	42699	21753	76192	10508
61	16309	20384	09491	91588	97720	89846	30376	76970	23063	35894
62	42587	37065	24526	72602	57589	98131	37292	05967	26002	51945
63	40177	98590	97161	41682	84533	67588	62036	49967	01990	72308
64	82309	76128	93965	26743	24141	04838	40254	26065	07938	76236
65	79788	68243	59732	04257	27084	14743	17520	95401	55811	76099
66	40538	79000	89559	25026	42274	23489	34502	75508	06059	86682
67	64016	73598	18609	73150	62463	33102	45205	87440	96767	67042
68	49767	12691	17903	93871	99721	79109	09425	26904	07419	76013
69	76974	55108	29795	08404	82684	00497	51126	79935	57450	55671
70	23854	08480	85983	96025	50117	64610	99425	62291	86943	21541
71	68973	70551	25098	78033	98573	79848	31778	29555	61446	23037
72	36444	93600	65350	14971	25325	00427	52073	64280	18847	24768
73	03003	87800	07391	11594	21196	00781	32550	57158	58887	73041
74	17540	26188	36647	78386	04558	61463	57842	90382	77019	24210
75	38916	55809	47982	41968	69760	79422	80154	91486	19180	15100
76	64288	19843	69122	42502	48508	28820	59933	72998	99942	10515
77	86809	51564	38040	39418	49915	19000	58050	16899	79952	57849
78	99800	99566	14742	05028	30033	94889	53381	23656	75787	59223
79	92345	31890	95712	08279	91794	94068	49337	88674	35355	12267
80	90363	65162	32245	82279	79256	80834	06088	99462	56705	06118
81	64437	32242	48431	04835	39070	59702	31508	60935	22390	52246
82	91714	53662	28373	34333	55791	74758	51144	18827	10704	76803
83	20902	17646	31391	31459	33315	03444	55743	74701	58851	27427
84	12217	86007	70371	52281	14510	76094	96579	54853	78339	20839
85	45177	02863	42307	53571	22532	74921	17735	42201	80540	54721
86	28325	90814	08804	52746	47913	54577	47525	77705	95330	21866
87	29019	28776	56116	54791	64604	08815	46049	71186	34650	14994
88	84979	81353	56219	67062	26146	82567	33122	14124	46240	92973
89	50371	26347	48513	63915	11158	25563	91915	18431	92978	11591
90	53422	06825	69711	67950	64716	18003	49581	45378	99878	61130
91	67453	35651	89316	41620	32048	70225	47597	33137	31443	51445
92	07294	85353	74819	23445	68237	07202	99515	62282	53809	26685
93	79544	00302	45338	16015	66613	88968	14595	63836	77716	79596
94	64144	85442	82060	46471	24162	39500	87351	36637	42833	71875
95	90919	11883	58318	00042	52402	28210	34075	33272	00840	73268
96	06670	57353	86275	92276	77591	46924	60839	55437	03183	13191
97	36634	93976	52062	83678	41256	60948	18685	48992	19462	96062
98	75101	72891	85745	67106	26010	62107	60885	37503	55461	71213
99	05112	71222	72654	51583	05228	62056	57390	42746	39272	96659

	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
50	32847	31282	03345	89593	69214	70381	78285	20054	91018	16742
51	16916	00041	30236	55023	14253	76582	12092	86533	92426	37655
52	66176	34037	21005	27137	03193	48970	64625	22394	39622	79085
53	46299	13335	12180	16861	38043	59292	62675	63631	37020	78195
54	22847	47839	45385	23289	47526	54098	45683	55849	51575	64689
55	41851	54160	92320	69936	34803	92479	33399	71160	64777	83378
56	28444	59497	91586	95917	68553	28639	06455	34174	11130	91994
57	47520	62378	98855	83174	13088	16561	68559	26679	06238	51254
58	34978	63271	13142	82681	05271	08822	06490	44984	49307	61717
59	37404	80416	69035	92980	49486	74378	75610	74976	70056	15478
60	32400	65482	52099	53676	74648	94148	65095	69597	52771	71551
61	89262	86332	51718	70663	11623	29834	79820	73002	84886	03591
62	86866	09127	98021	03871	27789	58444	44832	36505	40672	30180
63	90814	14833	08759	74645	05046	94056	99094	65091	32663	73040
64	19192	82756	20553	58446	55376	88914	75096	26119	83898	43816
65	77585	52593	56612	95766	10019	29531	73064	20953	53523	58136
66	23757	16364	05096	03192	62386	45389	85332	18877	55710	96459
67	45989	96257	23850	26216	23309	21526	07425	50254	19455	29315
68	92970	94243	07316	41467	64837	52406	25225	51553	31220	14032
69	74346	59596	40088	98176	17896	86900	20249	77753	19099	48885
70	87646	41309	27636	45153	29988	94770	07255	70908	05340	99751
71	50099	71038	45146	06146	55211	99429	43169	66259	97786	59180
72	10127	46900	64984	75348	04115	33624	68774	60013	35515	62556
73	67995	81977	18984	64091	02785	27762	42529	97144	80407	64524
74	26304	80217	84934	82657	69291	35397	98714	35104	08187	48109
75	81994	41070	56642	64091	31229	02595	13513	45148	78722	30144
76	59537	34662	79631	89403	65212	09975	06118	86197	58208	16162
77	51228	10937	62396	81460	47331	91403	95007	06047	16846	64809
78	31089	37995	29577	07828	42272	54016	21950	86192	99046	84864
79	38207	97938	93459	75174	79460	55436	57206	87644	21296	43393
80	88666	31142	09474	89712	63153	62333	42212	06140	42594	43671
81	53365	56134	67582	92557	89520	33452	05134	70628	27612	33738
82	89807	74530	38004	90102	11693	90257	05500	79920	62700	43325
83	18682	81038	85662	90915	91631	22223	91588	80774	07716	12548
84	63571	32579	63942	25371	09234	94592	98475	76884	37635	33608
85	68927	56492	67799	95398	77642	54913	91583	08421	81450	76229
86	56401	63186	39389	88798	31356	89235	97036	32341	33292	73757
87	24333	95603	02359	72942	46287	95382	08452	62862	97869	71775
88	17025	84202	95199	62272	06366	16175	97577	99304	41587	03686
89	02804	08253	52133	20224	68034	50865	57868	22343	55111	03607
90	08298	03879	20995	19850	73090	13191	18963	82244	78479	99121
91	59883	01785	82403	96062	03785	03488	12970	64896	38336	30030
92	46982	06682	62864	91837	74021	89094	39952	64158	79614	78235
93	31121	47266	07661	02051	67599	24471	69843	83696	71402	76287
94	97867	56641	63416	17577	30161	87320	37752	73276	48969	41915
95	57364	86746	08415	14621	49430	22311	15836	72492	49372	44103
96	09559	26263	69511	28064	75999	44540	13337	10918	79846	54809
97	53873	55571	00608	42661	91332	63956	74087	59008	47493	99581
98	35531	19162	86406	05299	77511	24311	57257	22826	77555	05941
99	28229	88629	25695	94932	30721	16197	78742	34974	97528	45447

TABLE A 3
CUMULATIVE NORMAL FREQUENCY DISTRIBUTION
(Area under the standard normal curve from 0 to Z)

[illegible]

TABLE A 4
THE DISTRIBUTION OF t^* (TWO-TAILED TESTS)

Degrees of Freedom	Probability of a Larger Value, Sign Ignored								
	0.500	0.400	0.200	0.100	0.050	0.025	0.010	0.005	0.001
1	1.000	1.376	3.078	6.314	12.706	25.452	63.657		
2	0.816	1.061	1.886	2.920	4.303	6.205	9.925	14.089	31.598
3	.765	0.978	1.638	2.353	3.182	4.176	5.841	7.453	12.941
4	.741	.941	1.533	2.132	2.776	3.495	4.604	5.598	8.610
5	.727	.920	1.476	2.015	2.571	3.163	4.032	4.773	6.859
6	.718	.906	1.440	1.943	2.447	2.969	3.707	4.317	5.959
7	.711	.896	1.415	1.895	2.365	2.841	3.499	4.029	5.405
8	.706	.889	1.397	1.860	2.306	2.752	3.355	3.832	5.041
9	.703	.883	1.383	1.833	2.262	2.685	3.250	3.690	4.781
10	.700	.879	1.372	1.812	2.228	2.634	3.169	3.581	4.587
11	.697	.876	1.363	1.796	2.201	2.593	3.106	3.497	4.437
12	.695	.873	1.356	1.782	2.179	2.560	3.055	3.428	4.318
13	.694	.870	1.350	1.771	2.160	2.533	3.012	3.372	4.221
14	.692	.868	1.345	1.761	2.145	2.510	2.977	3.326	4.140
15	.691	.866	1.341	1.753	2.131	2.490	2.947	3.286	4.073
16	.690	.865	1.337	1.746	2.120	2.473	2.921	3.252	4.015
17	.689	.863	1.333	1.740	2.110	2.458	2.898	3.222	3.965
18	.688	.862	1.330	1.734	2.101	2.445	2.878	3.197	3.922
19	.688	.861	1.328	1.729	2.093	2.433	2.861	3.174	3.883
20	.687	.860	1.325	1.725	2.086	2.423	2.845	3.153	3.850
21	.686	.859	1.323	1.721	2.080	2.414	2.831	3.135	3.819
22	.686	.858	1.321	1.717	2.074	2.406	2.819	3.119	3.792
23	.685	.858	1.319	1.714	2.069	2.398	2.807	3.104	3.767
24	.685	.857	1.318	1.711	2.064	2.391	2.797	3.090	3.745
25	.684	.856	1.316	1.708	2.060	2.385	2.787	3.078	3.725
26	.684	.856	1.315	1.706	2.056	2.379	2.779	3.067	3.707
27	.684	.855	1.314	1.703	2.052	2.373	2.771	3.056	3.690
28	.683	.855	1.313	1.701	2.048	2.368	2.763	3.047	3.674
29	.683	.854	1.311	1.699	2.045	2.364	2.756	3.038	3.659
30	.683	.854	1.310	1.697	2.042	2.360	2.750	3.030	3.646
35	.682	.852	1.306	1.690	2.030	2.342	2.724	2.996	3.591
40	.681	.851	1.303	1.684	2.021	2.329	2.704	2.971	3.551
45	.680	.850	1.301	1.680	2.014	2.319	2.690	2.952	3.520
50	.680	.849	1.299	1.676	2.008	2.310	2.678	2.937	3.496
55	.679	.849	1.297	1.673	2.004	2.304	2.669	2.925	3.476
60	.679	.848	1.296	1.671	2.000	2.299	2.660	2.915	3.460
70	.678	.847	1.294	1.667	1.994	2.290	2.648	2.899	3.435
80	.678	.847	1.293	1.665	1.989	2.284	2.638	2.887	3.416
90	.678	.846	1.291	1.662	1.986	2.279	2.631	2.878	3.402
100	.677	.846	1.290	1.661	1.982	2.276	2.625	2.871	3.390
120	.677	.845	1.289	1.658	1.980	2.270	2.617	2.860	3.373
∞	.6745	.8416	1.2816	1.6448	1.9600	2.2414	2.5758	2.8070	3.2905

* Parts of this table are reprinted by permission from R. A. Fisher's *Statistical Methods for Research Workers*, published by Oliver and Boyd, Edinburgh (1925-1950); from Maxine Merrington's "Table of Percentage Points of the t -Distribution," *Biometrika*, 32: 300 (1942); and from Bernard Ostle's *Statistics in Research*, Iowa State University Press (1954).

TABLE A 5
CUMULATIVE DISTRIBUTION OF CHI-SQUARE*

Degrees of Freedom	Probability of a Greater Value												
	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005
1	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00

TABLE A 5—(Continued)
CUMULATIVE DISTRIBUTION OF CHI-SQUARE*

Degrees of Freedom	Probability of a Greater Value												
	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.80	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17

* Condensed from table with 6 significant figures by Catherine M. Thompson, by permission of the Editor of *Biometrika*.

TABLE A 6
(i) TABLE FOR TESTING SKEWNESS
(One-tailed percentage points of the distribution of $\sqrt{b_1} = g_1 = m_3/m_2^{3/2}$)*

Size of Sample <i>n</i>	Percentage Points		Standard Deviation	Size of Sample <i>n</i>	Percentage Points		Standard Deviation
	5%	1%			5%	1%	
25	0.711	1.061	.4354	100	0.389	0.567	.2377
30	0.662	0.986	.4052	125	0.350	0.508	.2139
35	0.621	0.923	.3804	150	0.321	0.464	.1961
40	0.587	0.870	.3596	175	0.298	0.430	.1820
45	0.558	0.825	.3418	200	0.280	0.403	.1706
50	0.534	0.787	.3264				
				250	0.251	0.360	.1531
60	0.492	0.723	.3009	300	0.230	0.329	.1400
70	0.459	0.673	.2806	350	0.213	0.305	.1298
80	0.432	0.631	.2638	400	0.200	0.285	.1216
90	0.409	0.596	.2498	450	0.188	0.269	.1147
100	0.389	0.567	.2377	500	0.179	0.255	.1089

* Since the distribution of $\sqrt{b_1}$ is symmetrical about zero, the percentage points represent 10% and 2% two-tailed values. Reproduced from Table 34 B of *Tables for Statisticians and Biometricians*, Vol. 1, by permission of Dr. E. S. Pearson and the *Biometrika* Trustees.

TABLE A 6—(Continued)
(ii) TABLE FOR TESTING KURTOSIS
(Percentage points of the distribution of $b_2 = m_4/m_2^2$)*

Size of Sample <i>n</i>	Percentage Points				Size of Sample <i>n</i>	Percentage Points			
	Upper 1%	Upper 5%	Lower 5%	Lower 1%		Upper 1%	Upper 5%	Lower 5%	Lower 1%
50	4.88	3.99	2.15	1.95	600	3.54	3.34	2.70	2.60
75	4.59	3.87	2.27	2.08	650	3.52	3.33	2.71	2.61
100	4.39	3.77	2.35	2.18	700	3.50	3.31	2.72	2.62
125	4.24	3.71	2.40	2.24	750	3.48	3.30	2.73	2.64
150	4.13	3.65	2.45	2.29	800	3.46	3.29	2.74	2.65
					850	3.45	3.28	2.74	2.66
200	3.98	3.57	2.51	2.37	900	3.43	3.28	2.75	2.66
250	3.87	3.52	2.55	2.42	950	3.42	3.27	2.76	2.67
300	3.79	3.47	2.59	2.46	1000	3.41	3.26	2.76	2.68
350	3.72	3.44	2.62	2.50					
400	3.67	3.41	2.64	2.52	1200	3.37	3.24	2.78	2.71
450	3.63	3.39	2.66	2.55	1400	3.34	3.22	2.80	2.72
500	3.60	3.37	2.67	2.57	1600	3.32	3.21	2.81	2.74
550	3.57	3.35	2.69	2.58	1800	3.30	3.20	2.82	2.76
600	3.54	3.34	2.70	2.60	2000	3.28	3.18	2.83	2.77

* Reproduced from Table 34 C of *Tables for Statisticians and Biometricians*, by permission of Dr. E. S. Pearson and the *Biometrika* Trustees.

TABLE A 7

(1) SIGNIFICANCE LEVELS OF $t_w = (\bar{X} - \mu)/w$ IN NORMAL SAMPLES. TWO-TAILED TEST
DIVIDE P BY 2 FOR A ONE-TAILED TEST*

Size of Sample	Probability P			
	0.10	0.05	0.02	0.01
2	3.157	6.353	15.910	31.828
3	0.885	1.304	2.111	3.008
4	.529	0.717	1.023	1.316
5	.388	.507	0.685	0.843
6	.312	.399	.523	.628
7	.263	.333	.429	.507
8	.230	.288	.366	.429
9	.205	.255	.322	.374
10	.186	.230	.288	.333
11	.170	.210	.262	.302
12	.158	.194	.241	.277
13	.147	.181	.224	.256
14	.138	.170	.209	.239
15	.131	.160	.197	.224
16	.124	.151	.186	.212
17	.118	.144	.177	.201
18	.113	.137	.168	.191
19	.108	.131	.161	.182
20	.104	.126	.154	.175

* Taken from more extensive tables by permission of E. Lord and the Editor of *Biometrika*.

(Table A 7 continued overleaf)

TABLE A 7—(Continued)
(ii) SIGNIFICANCE LEVELS OF $(\bar{X}_1 - \bar{X}_2)/^{1/2}(w_1 + w_2)$ FOR TWO NORMAL
SAMPLES OF EQUAL SIZES.* TWO-TAILED TEST

Size of Sample	Probability P			
	0.10	0.05	0.02	0.01
2	2.322	3.427	5.553	7.916
3	0.974	1.272	1.715	2.093
4	.644	0.813	1.047	1.237
5	.493	.613	0.772	0.896
6	.405	.499	.621	.714
7	.347	.426	.525	.600
8	.306	.373	.459	.521
9	.275	.334	.409	.464
10	.250	.304	.371	.419
11	.233	.280	.340	.384
12	.214	.260	.315	.355
13	.201	.243	.294	.331
14	.189	.228	.276	.311
15	.179	.216	.261	.293
16	.170	.205	.247	.278
17	.162	.195	.236	.264
18	.155	.187	.225	.252
19	.149	.179	.216	.242
20	.143	.172	.207	.232

* From more extensive tables by permission of E. Lord and the Editor of *Biometrika*.

TABLE A 8
NUMBERS OF LIKE SIGNS REQUIRED FOR SIGNIFICANCE IN THE SIGN TEST,
WITH ACTUAL SIGNIFICANCE PROBABILITIES. TWO-TAILED TEST

No. of Pairs	Significance Level			No. of Pairs	Significance Level		
	1%	5%	10%		1%	5%	10%
5	0(.062)	13	1(.003)	2(.022)	3(.092)
6	0(.031)	0(.031)	14	1(.002)	2(.013)	3(.057)
7	0(.016)	0(.016)	15	2(.007)	3(.035)	3(.035)
8	0(.008)	0(.008)	1(.070)	16	2(.004)	3(.021)	4(.077)
9	0(.004)	1(.039)	1(.039)	17	2(.002)	4(.049)	4(.049)
10	0(.002)	1(.021)	1(.021)	18	3(.008)	4(.031)	5(.096)
11	0(.001)	1(.012)	2(.065)	19	3(.004)	4(.019)	5(.063)
12	1(.006)	2(.039)	2(.039)	20	3(.003)	5(.041)	5(.041)

TABLE A 9
SUM OF RANKS AT APPROXIMATE 5% AND 1% LEVELS OF P . * THESE NUMBERS
OR SMALLER INDICATE REJECTION. TWO-TAILED TEST

Number of Pairs	5% Level	1% Level
7	2(0.047)	0(0.016)
8	2(0.024)	0(0.008)
9	6(0.054)	2(0.009)
10	8(0.049)	3(0.010)
11	11(0.053)	5(0.009)
12	14(0.054)	7(0.009)
13	17(0.050)	10(0.010)
14	21(0.054)	13(0.011)
15	25(0.054)	16(0.010)
16	29(0.053)	19(0.009)

* The figures in parentheses are the actual significance probabilities. Adapted from the article by Wilcoxon (2, Chapter 5).

TABLE A 10
WILCOXON'S TWO-SAMPLE RANK TEST (THE MANN-WHITNEY TEST).
VALUES OF T AT TWO LEVELS
(These values or smaller cause rejection. Two-tailed test. Take $n_1 \leq n_2$ *)

0.05 Level of T														
$n_2 \downarrow$ $n_1 \rightarrow$	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4			10											
5		6	11	17										
6		7	12	18	26									
7		7	13	20	27	36								
8	3	8	14	21	29	38	49							
9	3	8	15	22	31	40	51	63						
10	3	9	15	23	32	42	53	65	78					
11	4	9	16	24	34	44	55	68	81	96				
12	4	10	17	26	35	46	58	71	85	99	115			
13	4	10	18	27	37	48	60	73	88	103	119	137		
14	4	11	19	28	38	50	63	76	91	106	123	141	160	
15	4	11	20	29	40	52	65	79	94	110	127	145	164	185
16	4	12	21	31	42	54	67	82	97	114	131	150	169	
17	5	12	21	32	43	56	70	84	100	117	135	154		
18	5	13	22	33	45	58	72	87	103	121	139			
19	5	13	23	34	46	60	74	90	107	124				
20	5	14	24	35	48	62	77	93	110					
21	6	14	25	37	50	64	79	95						
22	6	15	26	38	51	66	82							
23	6	15	27	39	53	68								
24	6	16	28	40	55									
25	6	16	28	42										
26	7	17	29											
27	7	17												
28	7													

(Table A 10 continued overleaf)

TABLE A 10—(Continued)

0.01 Level of T														
$n_2 \downarrow$ $n_1 \rightarrow$	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5				15										
6			10	16	23									
7			10	17	24	32								
8			11	17	25	34	43							
9		6	11	18	26	35	45	56						
10		6	12	19	27	37	47	58	71					
11		6	12	20	28	38	49	61	74	87				
12		7	13	21	30	40	51	63	76	90	106			
13		7	14	22	31	41	53	65	79	93	109	125		
14		7	14	22	32	43	54	67	81	96	112	129	147	
15		8	15	23	33	44	56	70	84	99	115	133	151	171
16		8	15	24	34	46	58	72	86	102	119	137	155	
17		8	16	25	36	47	60	74	89	105	122	140		
18		8	16	26	37	49	62	76	92	108	125			
19	3	9	17	27	38	50	64	78	94	111				
20	3	9	18	28	39	52	66	81	97					
21	3	9	18	29	40	53	68	83						
22	3	10	19	29	42	55	70							
23	3	10	19	30	43	57								
24	3	10	20	31	44									
25	3	11	20	32										
26	3	11	21											
27	4	11												
28	4													

* n_1 and n_2 are the numbers of cases in the two groups. If the groups are unequal in size, n_1 refers to the smaller.

Table is reprinted from White (12, Chapter 5) who extended the method of Wilcoxon.

TABLE A 11
CORRELATION COEFFICIENTS AT THE 5% AND 1% LEVELS OF SIGNIFICANCE

Degrees of Freedom	5%	1%	Degrees of Freedom	5%	1%
1	997	1 000	24	388	496
2	950	990	25	381	487
3	878	959	26	374	478
4	811	917	27	367	470
5	754	874	28	361	463
6	707	834	29	355	456
7	666	798	30	349	449
8	632	765	35	325	418
9	602	735	40	304	393
10	576	708	45	288	372
11	553	684	50	273	354
12	532	661	60	250	325
13	514	641	70	232	302
14	497	623	80	217	283
15	482	606	90	205	267
16	468	590	100	195	254
17	456	575	125	174	228
18	444	561	150	159	208
19	433	549	200	138	181
20	423	537	300	113	148
21	413	526	400	098	128
22	404	515	500	088	115
23	396	505	1,000	062	081

Portions of this table were taken from Table VA in *Statistical Methods for Research Workers* by permission of Professor R. A. Fisher and his publishers, Oliver and Boyd

TABLE A 12
TABLE OF $z = \frac{1}{2} \log_e (1 + r)/(1 - r)$ TO TRANSFORM THE CORRELATION COEFFICIENT

<i>r</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
.0	0.000	0.010	0.020	0.030	0.040	0.050	0.060	0.070	0.080	0.090
.1	.100	.110	.121	.131	.141	.151	.161	.172	.182	.192
.2	.203	.213	.224	.234	.245	.255	.266	.277	.288	.299
.3	.310	.321	.332	.343	.354	.365	.377	.388	.400	.412
.4	.424	.436	.448	.460	.472	.485	.497	.510	.523	.536
.5	.549	.563	.576	.590	.604	.618	.633	.648	.662	.678
.6	.693	.709	.725	.741	.758	.775	.793	.811	.829	.848
.7	.867	.887	.908	.929	.950	.973	.996	1.020	1.045	1.071
.8	1.099	1.127	1.157	1.188	1.221	1.256	1.293	1.333	1.376	1.422
<i>r</i>	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
.90	1.472	1.478	1.483	1.488	1.494	1.499	1.505	1.510	1.516	1.522
.91	1.528	1.533	1.539	1.545	1.551	1.557	1.564	1.570	1.576	1.583
.92	1.589	1.596	1.602	1.609	1.616	1.623	1.630	1.637	1.644	1.651
.93	1.658	1.666	1.673	1.681	1.689	1.697	1.705	1.713	1.721	1.730
.94	1.738	1.747	1.756	1.764	1.774	1.783	1.792	1.802	1.812	1.822
.95	1.832	1.842	1.853	1.863	1.874	1.886	1.897	1.909	1.921	1.933
.96	1.946	1.959	1.972	1.986	2.000	2.014	2.029	2.044	2.060	2.076
.97	2.092	2.109	2.127	2.146	2.165	2.185	2.205	2.227	2.249	2.273
.98	2.298	2.323	2.351	2.380	2.410	2.443	2.477	2.515	2.555	2.599
.99	2.646	2.700	2.759	2.826	2.903	2.994	3.106	3.250	3.453	3.800

TABLE A 13
TABLE OF r IN TERMS OF z^*

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.000	0.010	0.020	0.030	0.040	0.050	0.060	0.070	0.080	0.090
.1	.100	.110	.119	.129	.139	.149	.159	.168	.178	.187
.2	.197	.207	.216	.226	.236	.245	.254	.264	.273	.282
.3	.291	.300	.310	.319	.327	.336	.345	.354	.363	.371
.4	.380	.389	.397	.405	.414	.422	.430	.438	.446	.454
.5	.462	.470	.478	.485	.493	.500	.508	.515	.523	.530
.6	.537	.544	.551	.558	.565	.572	.578	.585	.592	.598
.7	.604	.611	.617	.623	.629	.635	.641	.647	.653	.658
.8	.664	.670	.675	.680	.686	.691	.696	.701	.706	.711
.9	.716	.721	.726	.731	.735	.740	.744	.74	.753	.757
1.0	.762	.766	.770	.774	.778	.782	.786	.790	.793	.797
1.1	.800	.804	.808	.811	.814	.818	.821	.824	.828	.831
1.2	.834	.837	.840	.843	.846	.848	.851	.854	.856	.859
1.3	.862	.864	.867	.869	.872	.874	.876	.879	.881	.883
1.4	.885	.888	.890	.892	.894	.896	.898	.900	.902	.903
1.5	.905	.907	.909	.910	.912	.914	.915	.917	.919	.920
1.6	.922	.923	.925	.926	.928	.929	.930	.932	.933	.934
1.7	.935	.937	.938	.939	.940	.941	.942	.944	.945	.946
1.8	.947	.948	.949	.950	.951	.952	.953	.954	.954	.955
1.9	.956	.957	.958	.959	.960	.960	.961	.962	.963	.963
2.0	.964	.965	.965	.966	.967	.967	.968	.969	.969	.970
2.1	.970	.971	.972	.972	.973	.973	.974	.974	.975	.975
2.2	.976	.976	.977	.977	.978	.978	.978	.979	.979	.980
2.3	.980	.980	.981	.981	.982	.982	.982	.983	.983	.983
2.4	.984	.984	.984	.985	.985	.985	.986	.986	.986	.986
2.5	.987	.987	.987	.987	.988	.988	.988	.988	.989	.989
2.6	.989	.989	.989	.990	.990	.990	.990	.990	.991	.991
2.7	.991	.991	.991	.992	.992	.992	.992	.992	.992	.992
2.8	.993	.993	.993	.993	.993	.993	.993	.994	.994	.994
2.9	.994	.994	.994	.994	.994	.995	.995	.995	.995	.995

* $r = (e^{2z} - 1)/(e^{2z} + 1)$.

TABLE A 14, Part I
5% (ROMAN TYPE) AND 1% (BOLD FACE TYPE) POINTS FOR THE DISTRIBUTION OF F

f_2		f_1 Degrees of Freedom (for greater mean square)																				f_2			
		1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75		100	200	500
1	161 4,052	200 4,999	216 5,403	225 5,625	230 5,764	234 5,859	237 5,928	239 5,981	241 6,022	242 6,056	243 6,082	244 6,106	245 6,142	246 6,169	248 6,208	249 6,234	250 6,261	251 6,286	252 6,302	253 6,323	253 6,334	254 6,352	254 6,361	254 6,366	
2	18.51 98.49	19.16 99.00	19.25 99.17	19.30 99.25	19.33 99.30	19.33 99.33	19.36 99.36	19.37 99.37	19.38 99.39	19.39 99.40	19.41 99.41	19.42 99.42	19.43 99.43	19.44 99.44	19.45 99.45	19.46 99.46	19.47 99.47	19.48 99.48	19.49 99.49	19.49 99.49	19.49 99.49	19.50 99.50	19.50 99.50	19.50 99.50	
3	10.13 34.12	9.55 30.82	9.28 29.46	9.12 28.71	9.01 28.24	8.94 27.91	8.88 27.67	8.84 27.49	8.81 27.34	8.78 27.23	8.76 27.13	8.74 27.05	8.71 26.92	8.69 26.83	8.66 26.69	8.64 26.60	8.62 26.50	8.60 26.41	8.58 26.35	8.57 26.27	8.56 26.23	8.54 26.18	8.54 26.14	8.53 26.12	
4	7.71 21.20	6.94 18.00	6.59 16.69	6.39 15.98	6.26 15.52	6.16 15.21	6.09 14.98	6.04 14.80	6.00 14.66	5.96 14.54	5.93 14.45	5.91 14.37	5.87 14.24	5.84 14.15	5.80 14.02	5.77 13.93	5.74 13.83	5.71 13.74	5.70 13.69	5.68 13.61	5.66 13.57	5.65 13.52	5.64 13.48	5.63 13.46	
5	6.61 16.26	5.79 13.27	5.41 12.06	5.19 11.39	5.05 10.97	4.95 10.67	4.88 10.45	4.82 10.29	4.78 10.15	4.74 10.05	4.70 9.96	4.68 9.89	4.64 9.77	4.60 9.68	4.56 9.55	4.53 9.47	4.50 9.38	4.46 9.29	4.44 9.24	4.42 9.17	4.40 9.13	4.37 9.07	4.36 9.04	4.36 9.02	
6	5.99 13.74	5.14 10.92	4.76 9.78	4.53 9.15	4.39 8.75	4.28 8.47	4.21 8.26	4.15 8.10	4.10 7.98	4.06 7.87	4.03 7.79	4.00 7.72	3.96 7.60	3.92 7.52	3.87 7.39	3.84 7.31	3.81 7.23	3.77 7.14	3.75 7.09	3.72 7.02	3.71 6.99	3.69 6.94	3.68 6.90	3.67 6.88	
7	5.59 12.25	4.74 9.55	4.35 8.45	4.12 7.85	3.97 7.46	3.87 7.19	3.79 7.00	3.73 6.84	3.68 6.71	3.63 6.62	3.60 6.54	3.57 6.47	3.52 6.35	3.49 6.27	3.44 6.15	3.41 6.07	3.38 5.98	3.34 5.90	3.32 5.85	3.29 5.78	3.28 5.75	3.25 5.70	3.24 5.67	3.23 5.65	
8	5.32 11.26	4.46 8.65	4.07 7.59	3.84 7.01	3.69 6.63	3.58 6.37	3.50 6.19	3.44 6.03	3.39 5.91	3.34 5.82	3.31 5.74	3.28 5.67	3.23 5.56	3.20 5.48	3.15 5.36	3.12 5.28	3.08 5.20	3.05 5.11	3.03 5.06	3.00 5.00	2.98 4.96	2.96 4.91	2.94 4.88	2.93 4.86	
9	5.12 10.56	4.26 8.02	3.86 6.99	3.63 6.42	3.48 6.06	3.37 5.80	3.29 5.62	3.23 5.47	3.18 5.35	3.13 5.26	3.10 5.18	3.07 5.11	3.02 5.00	2.98 4.92	2.93 4.80	2.90 4.73	2.86 4.64	2.82 4.56	2.80 4.45	2.77 4.41	2.76 4.36	2.73 4.33	2.72 4.31	2.71 4.31	
10	4.96 10.04	4.10 7.56	3.71 6.55	3.48 5.99	3.33 5.64	3.22 5.39	3.14 5.21	3.07 5.06	3.02 4.95	2.97 4.85	2.94 4.78	2.91 4.71	2.86 4.60	2.82 4.52	2.77 4.41	2.74 4.33	2.70 4.25	2.67 4.17	2.64 4.12	2.61 4.05	2.59 3.96	2.55 3.93	2.54 3.91	2.54 3.91	
11	4.84 9.65	3.98 7.20	3.59 6.22	3.36 5.67	3.20 5.32	3.09 5.07	3.01 4.88	2.95 4.74	2.90 4.63	2.86 4.54	2.82 4.46	2.79 4.40	2.74 4.29	2.70 4.21	2.65 4.10	2.61 4.02	2.57 3.94	2.53 3.86	2.50 3.80	2.47 3.74	2.45 3.70	2.42 3.66	2.41 3.62	2.40 3.60	
12	4.75 9.33	3.88 6.93	3.49 5.95	3.26 5.41	3.11 5.06	3.00 4.82	2.92 4.65	2.85 4.50	2.80 4.39	2.76 4.30	2.72 4.22	2.69 4.16	2.64 4.05	2.60 3.98	2.54 3.86	2.50 3.78	2.46 3.70	2.42 3.61	2.36 3.56	2.32 3.49	2.32 3.46	2.31 3.38	2.30 3.36	2.30 3.36	
13	4.67 9.07	3.80 6.70	3.41 5.74	3.18 5.20	3.02 4.86	2.92 4.62	2.84 4.44	2.77 4.30	2.72 4.19	2.67 4.10	2.63 4.02	2.60 3.96	2.55 3.85	2.51 3.78	2.46 3.67	2.42 3.59	2.38 3.51	2.34 3.42	2.32 3.37	2.28 3.30	2.26 3.27	2.24 3.24	2.22 3.21	2.22 3.18	2.21 3.16

TABLE A 14, Part I—(Continued)

f_2	f_1 Degrees of Freedom (for greater mean square)																								f_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13	14
	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.70	3.62	3.51	3.43	3.34	3.26	3.21	3.14	3.11	3.06	3.02	3.00	15
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07	15
	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.56	3.48	3.36	3.29	3.20	3.12	3.07	3.00	2.97	2.92	2.89	2.87	16
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01	16
	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.61	3.55	3.45	3.37	3.25	3.18	3.10	3.01	2.96	2.98	2.86	2.80	2.77	2.75	17
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96	17
	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45	3.35	3.27	3.16	3.08	3.00	2.92	2.86	2.79	2.76	2.70	2.67	2.65	18
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92	18
	8.28	6.01	5.09	4.58	4.25	4.01	3.85	3.71	3.60	3.51	3.44	3.37	3.27	3.19	3.07	3.00	2.91	2.83	2.78	2.71	2.68	2.62	2.59	2.57	19
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88	19
	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.19	3.12	3.00	2.92	2.84	2.76	2.70	2.63	2.60	2.54	2.51	2.49	20
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84	20
	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.30	3.23	3.13	3.05	2.94	2.86	2.77	2.69	2.63	2.56	2.53	2.47	2.44	2.42	21
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81	21
	8.02	5.78	4.87	4.37	4.04	3.81	3.65	3.51	3.40	3.31	3.24	3.17	3.07	2.99	2.88	2.80	2.72	2.63	2.58	2.51	2.47	2.42	2.38	2.36	22
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78	22
	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.02	2.94	2.83	2.75	2.67	2.58	2.53	2.46	2.42	2.37	2.33	2.31	23
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76	23
	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.97	2.89	2.78	2.70	2.62	2.53	2.48	2.41	2.37	2.32	2.28	2.26	24
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73	24
	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09	3.03	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.36	2.33	2.27	2.23	2.21	25
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71	25
	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.21	3.13	3.05	2.99	2.89	2.81	2.70	2.62	2.54	2.45	2.40	2.32	2.29	2.23	2.19	2.17	26
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69	26
	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	3.02	2.96	2.86	2.77	2.66	2.58	2.50	2.41	2.36	2.28	2.25	2.19	2.15	2.13	

The function, $F = e$ with exponent $2z$, is computed in part from Fisher's table VI (7). Additional entries are by interpolation, mostly graphical.

TABLE A 14. Part I—(Continued)

		f_1 Degrees of Freedom (for greater mean square)																				f_2			
f_1		1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
27	4.21 7.68	3.35 5.49	2.96 4.60	2.73 4.11	2.57 3.79	2.46 3.56	2.37 3.39	2.30 3.26	2.25 3.14	2.20 3.06	2.16 2.98	2.13 2.93	2.08 2.83	2.03 2.74	1.97 2.63	1.93 2.55	1.88 2.47	1.84 2.38	1.80 2.33	1.76 2.25	1.74 2.21	1.71 2.16	1.68 2.12	1.67 2.10	
28	4.20 7.64	3.34 5.45	2.95 4.57	2.71 4.07	2.56 3.76	2.44 3.53	2.36 3.36	2.29 3.23	2.24 3.11	2.19 3.03	2.15 2.95	2.12 2.90	2.06 2.80	2.02 2.71	1.96 2.60	1.91 2.52	1.87 2.44	1.81 2.35	1.78 2.30	1.75 2.22	1.72 2.18	1.69 2.13	1.67 2.09	1.65 2.06	
29	4.18 7.60	3.33 5.42	2.93 4.54	2.70 4.04	2.54 3.73	2.43 3.50	2.35 3.33	2.28 3.20	2.22 3.08	2.18 3.00	2.14 2.92	2.10 2.87	2.05 2.77	2.00 2.68	1.94 2.57	1.90 2.49	1.85 2.41	1.80 2.32	1.77 2.27	1.73 2.19	1.71 2.15	1.68 2.10	1.65 2.06	1.64 2.03	
30	4.17 7.56	3.32 5.39	2.92 4.51	2.69 4.02	2.53 3.70	2.42 3.47	2.34 3.30	2.27 3.17	2.21 3.06	2.16 2.98	2.12 2.90	2.09 2.84	2.04 2.74	1.99 2.66	1.93 2.55	1.89 2.47	1.84 2.38	1.79 2.29	1.76 2.24	1.72 2.16	1.69 2.13	1.66 2.07	1.64 2.03	1.62 2.01	
32	4.15 7.50	3.30 5.34	2.90 4.46	2.67 3.97	2.51 3.66	2.40 3.42	2.32 3.25	2.25 3.12	2.19 3.01	2.14 2.94	2.10 2.86	2.07 2.80	2.02 2.70	1.97 2.62	1.91 2.51	1.86 2.42	1.82 2.34	1.76 2.25	1.74 2.20	1.69 2.12	1.67 2.08	1.64 2.02	1.61 1.98	1.59 1.96	
34	4.13 7.44	3.28 5.29	2.88 4.42	2.65 3.93	2.49 3.61	2.38 3.38	2.30 3.21	2.23 3.08	2.17 2.97	2.12 2.89	2.08 2.82	2.05 2.76	2.00 2.66	1.95 2.58	1.89 2.47	1.84 2.38	1.80 2.30	1.74 2.21	1.71 2.15	1.67 2.08	1.64 2.04	1.61 1.98	1.59 1.94	1.57 1.91	
36	4.11 7.39	3.26 5.25	2.86 4.38	2.63 3.89	2.48 3.58	2.36 3.35	2.28 3.18	2.21 3.04	2.15 2.94	2.10 2.86	2.06 2.78	2.03 2.72	1.98 2.62	1.93 2.54	1.87 2.43	1.82 2.35	1.78 2.26	1.72 2.17	1.69 2.12	1.65 2.04	1.62 2.00	1.59 1.94	1.56 1.90	1.55 1.87	
38	4.10 7.35	3.25 5.21	2.85 4.34	2.62 3.86	2.46 3.54	2.35 3.32	2.26 3.15	2.19 3.02	2.14 2.91	2.09 2.82	2.05 2.75	2.02 2.69	1.96 2.59	1.92 2.51	1.85 2.40	1.80 2.32	1.76 2.22	1.71 2.14	1.67 2.08	1.63 2.00	1.60 1.97	1.57 1.90	1.54 1.86	1.53 1.84	
40	4.08 7.31	3.23 5.18	2.84 4.31	2.61 3.83	2.45 3.51	2.34 3.29	2.25 3.12	2.18 2.99	2.12 2.88	2.07 2.80	2.04 2.73	2.00 2.66	1.95 2.56	1.90 2.49	1.84 2.37	1.79 2.29	1.74 2.20	1.69 2.11	1.66 2.05	1.61 1.97	1.59 1.94	1.55 1.88	1.53 1.84	1.51 1.81	
42	4.07 7.27	3.22 5.15	2.83 4.29	2.59 3.80	2.44 3.49	2.32 3.26	2.24 3.10	2.17 2.96	2.11 2.86	2.06 2.77	2.02 2.70	1.99 2.64	1.94 2.54	1.89 2.46	1.82 2.35	1.78 2.26	1.73 2.17	1.68 2.08	1.64 2.02	1.60 1.94	1.57 1.91	1.54 1.85	1.51 1.80	1.49 1.78	
44	4.06 7.24	3.21 5.12	2.82 4.26	2.58 3.78	2.43 3.46	2.31 3.24	2.23 3.07	2.16 2.94	2.10 2.84	2.05 2.75	2.01 2.68	1.98 2.52	1.92 2.44	1.88 2.32	1.81 2.24	1.76 2.22	1.72 2.15	1.66 2.06	1.63 1.92	1.58 1.88	1.56 1.91	1.52 1.82	1.50 1.78	1.48 1.75	
46	4.05 7.21	3.20 5.10	2.81 4.24	2.57 3.76	2.42 3.44	2.30 3.22	2.22 3.05	2.14 2.92	2.09 2.82	2.04 2.73	2.00 2.66	1.97 2.50	1.91 2.42	1.87 2.30	1.80 2.22	1.75 2.20	1.71 2.13	1.65 2.04	1.62 1.98	1.57 1.90	1.54 1.86	1.51 1.80	1.48 1.76	1.46 1.72	
48	4.04 7.19	3.19 5.08	2.80 4.22	2.56 3.74	2.41 3.42	2.30 3.20	2.21 3.04	2.14 2.90	2.08 2.80	2.03 2.71	1.99 2.64	1.96 2.58	1.90 2.48	1.86 2.40	1.79 2.28	1.74 2.20	1.70 2.11	1.64 2.02	1.61 1.96	1.56 1.88	1.53 1.84	1.50 1.78	1.47 1.73	1.45 1.70	

TABLE A 14, Part I—(Continued)

f_2	f_1 Degrees of Freedom (for greater mean square)																				f_2		
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	∞			
50	4.03 7.17	3.18 5.06	2.79 4.20	2.56 3.72	2.40 3.41	2.29 3.18	2.20 3.02	2.13 2.88	2.07 2.78	2.02 2.70	1.98 2.62	1.95 2.56	1.90 2.46	1.85 2.39	1.85 2.26	1.78 2.18	1.69 2.10	1.63 2.00	1.60 1.94	1.55 1.86	1.52 1.82	1.48 1.71	1.46 1.68
55	4.02 7.12	3.17 5.01	2.78 4.16	2.54 3.68	2.38 3.37	2.27 3.15	2.18 2.98	2.11 2.85	2.05 2.75	2.00 2.66	1.97 2.59	1.93 2.53	1.88 2.43	1.83 2.35	1.76 2.23	1.72 2.15	1.67 2.06	1.61 1.96	1.58 1.90	1.52 1.82	1.50 1.77	1.46 1.66	1.43 1.64
60	4.00 7.08	3.15 4.98	2.76 4.13	2.52 3.65	2.37 3.34	2.25 3.12	2.17 2.95	2.10 2.82	2.04 2.72	1.99 2.63	1.95 2.56	1.92 2.50	1.86 2.40	1.81 2.32	1.75 2.20	1.70 2.12	1.65 2.03	1.59 1.93	1.56 1.87	1.50 1.79	1.48 1.74	1.44 1.68	1.41 1.60
65	3.99 7.04	3.14 4.95	2.75 4.10	2.51 3.62	2.36 3.31	2.24 3.09	2.15 2.93	2.08 2.79	2.02 2.70	1.98 2.61	1.94 2.54	1.90 2.47	1.85 2.37	1.80 2.30	1.73 2.18	1.68 2.09	1.63 2.00	1.57 1.90	1.54 1.84	1.49 1.76	1.46 1.71	1.42 1.64	1.39 1.60
70	3.98 7.01	3.13 4.92	2.74 4.08	2.50 3.60	2.35 3.29	2.23 3.07	2.14 2.91	2.07 2.77	2.01 2.67	1.97 2.59	1.93 2.51	1.89 2.45	1.84 2.35	1.79 2.28	1.72 2.15	1.67 2.07	1.62 1.98	1.56 1.88	1.53 1.82	1.47 1.74	1.45 1.69	1.40 1.62	1.37 1.56
80	3.96 6.96	3.11 4.88	2.72 4.04	2.48 3.56	2.33 3.25	2.21 3.04	2.12 2.87	2.05 2.74	1.99 2.64	1.95 2.55	1.91 2.48	1.88 2.41	1.82 2.32	1.77 2.24	1.70 2.11	1.65 2.03	1.60 1.94	1.54 1.84	1.51 1.78	1.45 1.70	1.42 1.65	1.38 1.57	1.35 1.52
100	3.94 6.90	3.09 4.82	2.70 3.98	2.46 3.51	2.30 3.20	2.19 2.99	2.10 2.82	2.03 2.69	1.97 2.59	1.92 2.51	1.88 2.43	1.85 2.36	1.79 2.26	1.75 2.19	1.68 2.06	1.63 1.98	1.57 1.89	1.51 1.79	1.48 1.73	1.42 1.64	1.39 1.54	1.30 1.46	1.28 1.43
125	3.92 6.84	3.07 4.78	2.68 3.94	2.44 3.47	2.29 3.17	2.17 2.95	2.08 2.79	2.01 2.65	1.95 2.56	1.90 2.47	1.86 2.40	1.83 2.33	1.77 2.23	1.72 2.15	1.65 2.03	1.60 1.94	1.55 1.85	1.49 1.75	1.45 1.68	1.39 1.59	1.36 1.46	1.31 1.40	1.27 1.37
150	3.91 6.81	3.06 4.75	2.67 3.91	2.43 3.44	2.27 3.14	2.16 2.92	2.07 2.76	2.00 2.62	1.94 2.53	1.89 2.44	1.85 2.36	1.82 2.30	1.76 2.20	1.71 2.12	1.64 2.00	1.59 1.91	1.54 1.83	1.47 1.72	1.44 1.66	1.37 1.51	1.34 1.43	1.29 1.37	1.25 1.33
200	3.89 6.76	3.04 4.71	2.65 3.88	2.41 3.41	2.26 3.11	2.14 2.90	2.05 2.73	1.98 2.60	1.92 2.50	1.87 2.41	1.83 2.28	1.80 2.27	1.74 2.17	1.69 2.09	1.62 1.97	1.57 1.88	1.52 1.79	1.45 1.69	1.42 1.62	1.35 1.53	1.32 1.48	1.22 1.39	1.19 1.28
400	3.86 6.70	3.02 4.66	2.62 3.83	2.39 3.36	2.23 3.06	2.12 2.85	2.03 2.69	1.96 2.55	1.90 2.46	1.85 2.37	1.81 2.29	1.78 2.23	1.72 2.12	1.67 2.04	1.60 1.92	1.54 1.84	1.49 1.74	1.42 1.64	1.38 1.57	1.32 1.47	1.28 1.42	1.22 1.32	1.16 1.24
1000	3.85 6.66	3.00 4.62	2.61 3.80	2.38 3.34	2.22 3.04	2.10 2.82	2.02 2.66	1.95 2.53	1.89 2.43	1.84 2.34	1.80 2.26	1.76 2.20	1.70 2.09	1.65 2.01	1.58 1.89	1.53 1.81	1.47 1.71	1.41 1.61	1.36 1.54	1.30 1.44	1.26 1.36	1.19 1.28	1.13 1.18
∞	3.84 6.64	2.99 4.60	2.60 3.78	2.37 3.32	2.21 3.02	2.09 2.80	2.01 2.64	1.94 2.51	1.88 2.41	1.83 2.32	1.79 2.24	1.75 2.18	1.69 2.07	1.64 1.99	1.57 1.87	1.52 1.79	1.46 1.69	1.40 1.59	1.35 1.52	1.28 1.41	1.24 1.36	1.17 1.25	1.11 1.15

TABLE A 14. PART II
25%, 10%, 2.5%, AND 0.5% POINTS FOR THE DISTRIBUTION OF F^*

f_2	P	f_1 Degrees of Freedom (for greater mean square)																	120	∞
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60		
1	0.250	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.41	9.49	9.58	9.63	9.67	9.71	9.76	9.80	9.85
	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.20	60.70	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
	.025	648	800	864	900	922	937	948	957	963	969	977	985	993	997	1,001	1,006	1,010	1,014	1,018
	.005	16,211	20,000	21,615	22,500	23,056	23,437	23,715	23,925	24,091	24,224	24,426	24,630	24,836	24,940	25,044	25,148	25,253	25,359	25,465
2	.250	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.41	3.43	3.43	3.44	3.45	3.46	3.47	3.48
	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
	.025	38.51	39.00	39.16	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.42	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
	.005	198	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	200
3	.250	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.46	2.46	2.46	2.46	2.47	2.47	2.47	2.47
	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
	.025	17.44	16.04	15.44	15.10	14.88	14.74	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
	.005	55.55	49.80	47.47	46.20	45.39	44.84	44.43	44.13	43.88	43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83
4	.250	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
	.100	4.94	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
	.005	31.33	26.28	24.26	23.16	22.46	21.98	21.62	21.35	21.14	20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32
5	.250	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.88	1.88	1.88	1.88	1.87	1.87	1.87
	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
	.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.14
6	.250	1.62	1.76	1.78	1.79	1.79	1.79	1.78	1.78	1.77	1.77	1.77	1.76	1.76	1.75	1.75	1.75	1.74	1.74	1.74
	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
	.025	9.01	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
	.005	18.64	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88
7	.250	1.57	1.70	1.72	1.72	1.71	1.71	1.71	1.70	1.69	1.69	1.68	1.68	1.67	1.67	1.66	1.66	1.65	1.65	1.65
	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
	.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.64	7.53	7.42	7.31	7.19	7.08
8	.250	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.64	1.63	1.62	1.62	1.62	1.60	1.60	1.59	1.59	1.58	1.58
	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.43	4.36	4.30	4.24	4.14	4.04	3.95	3.89	3.84	3.78	3.73	3.67	3.61
	.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95

TABLE A 14, Part II—(Continued)
25%, 10%, 2.5%, AND 0.5% POINTS FOR THE DISTRIBUTION OF F^*

f_1	P	f_2 Degrees of Freedom (for greater mean square)																	∞
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
9	0.250	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.54	1.53
	.100	3.16	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18
	.025	7.21	5.71	5.08	4.62	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.33
	.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30
10	.250	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.53	1.52	1.52	1.51	1.51	1.50	1.48
	.100	3.28	2.92	2.73	2.61	2.52	2.46	2.42	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08
	.025	6.94	5.42	4.83	4.47	4.34	4.24	4.07	3.95	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.06
	.005	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75
11	.250	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.45
	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.33	2.28	2.21	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	1.97
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.74	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.83
	.005	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.24	5.05	4.86	4.76	4.65	4.55	4.44	4.34
12	.250	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.45	1.44	1.42
	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79
	.005	11.73	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01
13	.250	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.40
	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.85
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66
	.005	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76
14	.250	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.41	1.40	1.38
	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55
	.005	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55
15	.250	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.43	1.41	1.41	1.40	1.39	1.38	1.36
	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79
	.025	6.20	4.76	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.58	2.52	2.46
	.005	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37
16	.250	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35
	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38
	.005	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22
17	.250	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33
	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	1.96	1.91	1.86	1.82	1.80	1.78	1.75	1.72	1.69
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32
	.005	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10

TABLE A. 14, Part II—(Continued)
25%, 10%, 5%, AND 0.5% POINTS FOR THE DISTRIBUTION OF F^*

f_1	P	f_2 Degrees of Freedom (for greater mean square)																	∞
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	50	
18	0.250	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.32
	0.100	1.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.86	1.81	1.78	1.75	1.72	1.69
	0.05	0.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.44	2.35	2.28	2.22	2.19
	0.005	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99
19	0.250	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.33	1.32
	0.100	0.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67
	0.05	0.95	4.52	3.91	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.43	2.39	2.33	2.27	2.20
	0.005	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89
20	0.250	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31
	0.100	0.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64
	0.05	0.94	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16
	0.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81
21	0.250	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30
	0.100	0.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.88	1.83	1.78	1.75	1.72	1.69	1.66	1.62
	0.05	0.93	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11
	0.005	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73
22	0.250	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.37	1.36	1.34	1.33	1.32	1.31	1.30	1.29
	0.100	0.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60
	0.05	0.92	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08
	0.005	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66
23	0.250	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28
	0.100	0.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59
	0.05	0.91	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04
	0.005	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60
24	0.250	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.36	1.35	1.33	1.32	1.31	1.30	1.29	1.28
	0.100	0.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57
	0.05	0.90	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01
	0.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55
25	0.250	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.32	1.31	1.29	1.28	1.25
	0.100	0.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56
	0.05	0.89	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98
	0.005	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.37	3.20	3.02	2.92	2.82	2.72	2.61	2.50
26	0.250	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.35	1.34	1.32	1.31	1.30	1.29	1.28	1.26
	0.100	0.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54
	0.05	0.88	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95
	0.005	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45

TABLE A 14, Part II—(Continued)
2.5%, 10%, 2.5%, AND 0.5% POINTS FOR THE DISTRIBUTION OF F^*

f_2		f_1 Degrees of Freedom (for greater mean square)																			∞
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120		
27	0.250	1.38	1.46	1.45	1.43	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.33	1.32	1.31	1.30	1.28	1.27	1.26	1.24	
	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49	
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85	
	.005	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29	
28	.250	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.31	1.30	1.29	1.28	1.27	1.25	1.24	
	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48	
	.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83	
	.005	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25	
29	.250	1.38	1.45	1.45	1.43	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.32	1.31	1.30	1.29	1.27	1.26	1.25	1.23	
	.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47	
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81	
	.005	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.21	
30	.250	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.34	1.32	1.30	1.29	1.28	1.27	1.26	1.24	1.23	
	.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46	
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79	
	.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18	
40	.250	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.31	1.30	1.28	1.26	1.25	1.24	1.22	1.21	1.19	
	.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.60	1.57	1.54	1.51	1.47	1.42	1.38	
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64	
	.005	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	1.93	
60	.250	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.27	1.25	1.25	1.24	1.22	1.21	1.17	1.15	
	.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29	
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48	
	.005	8.49	5.80	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.69	
120	.250	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.26	1.24	1.22	1.21	1.19	1.18	1.16	1.13	1.10	
	.100	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.54	1.48	1.45	1.41	1.37	1.32	1.26	1.19	
	.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31	
	.005	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.43	
∞	.250	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.22	1.19	1.18	1.16	1.14	1.12	1.08	1.00	
	.100	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00	
	.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00	
	.005	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.00	

* Reprinted from "Tables of percentage points of the inverted beta (F) distribution," by Maxine Merrington and Catherine M. Thompson. *Biometrika*, 33:73 (1943) by permission of the authors and the editor.

TABLE A 15
UPPER 5% PERCENTAGE POINTS, Q , IN THE STUDENTIZED RANGE*

Degrees of Freedom, ν	Number of Treatments, a																			
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	18.0	27.0	32.8	37.2	40.5	43.1	45.4	47.3	49.1	50.6	51.9	53.2	54.3	55.4	56.3	57.2	58.0	58.8	59.6	
2	6.09	8.33	9.80	10.89	11.73	12.43	13.03	13.54	13.99	14.39	14.75	15.08	15.38	15.65	15.91	16.14	16.36	16.57	16.77	
3	4.50	5.91	6.83	7.51	8.04	8.47	8.85	9.18	9.46	9.72	9.95	10.16	10.35	10.52	10.69	10.84	10.98	11.12	11.24	
4	3.93	5.04	5.76	6.29	6.71	7.06	7.35	7.60	7.83	8.03	8.21	8.37	8.52	8.67	8.80	8.92	9.03	9.14	9.24	
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21	
6	3.46	4.34	4.90	5.31	5.63	5.89	6.12	6.32	6.49	6.65	6.79	6.92	7.04	7.14	7.24	7.34	7.43	7.51	7.59	
7	3.34	4.16	4.68	5.06	5.35	5.59	5.80	5.99	6.15	6.29	6.42	6.54	6.65	6.75	6.84	6.93	7.01	7.08	7.16	
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87	
9	3.20	3.95	4.42	4.76	5.02	5.24	5.43	5.60	5.74	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.65	
10	3.15	3.88	4.33	4.66	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.12	6.20	6.27	6.34	6.41	6.47	
11	3.11	3.82	4.26	4.58	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.98	6.06	6.14	6.20	6.27	6.33	
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40	5.51	5.61	5.71	5.80	5.88	5.95	6.02	6.09	6.15	6.21	
13	3.06	3.73	4.15	4.46	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79	5.86	5.93	6.00	6.06	6.11	
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.56	5.64	5.72	5.79	5.86	5.92	5.98	6.03	
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.57	5.65	5.72	5.79	5.85	5.91	5.96	
16	3.00	3.65	4.05	4.34	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.90	
17	2.98	3.62	4.02	4.31	4.52	4.70	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.55	5.61	5.68	5.74	5.79	5.84	
18	2.97	3.61	4.00	4.28	4.49	4.67	4.83	4.96	5.07	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79	
19	2.96	3.59	3.98	4.26	4.47	4.64	4.79	4.92	5.04	5.14	5.23	5.32	5.39	5.46	5.53	5.59	5.65	5.70	5.75	
20	2.95	3.58	3.96	4.24	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43	5.50	5.56	5.61	5.66	5.71	
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32	5.38	5.44	5.50	5.55	5.59	
30	2.89	3.48	3.84	4.11	4.30	4.46	4.60	4.72	4.83	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.48	
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.74	4.82	4.90	4.98	5.05	5.11	5.17	5.22	5.27	5.32	5.36	
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00	5.06	5.11	5.15	5.20	5.24	
120	2.80	3.36	3.69	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71	4.78	4.84	4.90	4.95	5.00	5.04	5.09	5.13	
∞	2.77	3.32	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.68	4.74	4.80	4.84	4.89	4.93	4.97	5.01	

* Reprinted from *Biometrika*, 39:192 (1952) by permission of the author, Joyce M. May, and the editor.

TABLE A 16
 ANGLES CORRESPONDING TO PERCENTAGES, $\text{ANGLE} = \text{ARCSIN} \sqrt{\text{PERCENTAGE}}$,
 AS GIVEN BY C. I. BLISS*

%	0	1	2	3	4	5	6	7	8	9
0.0	0	0.57	0.81	0.99	1.15	1.28	1.40	1.52	1.62	1.72
0.1	1.81	1.90	1.99	2.07	2.14	2.22	2.29	2.36	2.43	2.50
0.2	2.56	2.63	2.69	2.75	2.81	2.87	2.92	2.98	3.03	3.09
0.3	3.14	3.19	3.24	3.29	3.34	3.39	3.44	3.49	3.53	3.58
0.4	3.63	3.67	3.72	3.76	3.80	3.85	3.89	3.93	3.97	4.01
0.5	4.05	4.09	4.13	4.17	4.21	4.25	4.29	4.33	4.37	4.40
0.6	4.44	4.48	4.52	4.55	4.59	4.62	4.66	4.69	4.73	4.76
0.7	4.80	4.83	4.87	4.90	4.93	4.97	5.00	5.03	5.07	5.10
0.8	5.13	5.16	5.20	5.23	5.26	5.29	5.32	5.35	5.38	5.41
0.9	5.44	5.47	5.50	5.53	5.56	5.59	5.62	5.65	5.68	5.71
1	5.74	6.02	6.29	6.55	6.80	7.04	7.27	7.49	7.71	7.92
2	8.13	8.33	8.53	8.72	8.91	9.10	9.28	9.46	9.63	9.81
3	9.98	10.14	10.31	10.47	10.63	10.78	10.94	11.09	11.24	11.39
4	11.54	11.68	11.83	11.97	12.11	12.25	12.39	12.52	12.66	12.79
5	12.92	13.05	13.18	13.31	13.44	13.56	13.69	13.81	13.94	14.06
6	14.18	14.30	14.42	14.54	14.65	14.77	14.89	15.00	15.12	15.23
7	15.34	15.45	15.56	15.68	15.79	15.89	16.00	16.11	16.22	16.32
8	16.43	16.54	16.64	16.74	16.85	16.95	17.05	17.16	17.26	17.36
9	17.46	17.56	17.66	17.76	17.85	17.95	18.05	18.15	18.24	18.34
10	18.44	18.53	18.63	18.72	18.81	18.91	19.00	19.09	19.19	19.28
11	19.37	19.46	19.55	19.64	19.73	19.82	19.91	20.00	20.09	20.18
12	20.27	20.36	20.44	20.53	20.62	20.70	20.79	20.88	20.96	21.05
13	21.13	21.22	21.30	21.39	21.47	21.56	21.64	21.72	21.81	21.89
14	21.97	22.06	22.14	22.22	22.30	22.38	22.46	22.55	22.63	22.71
15	22.79	22.87	22.95	23.03	23.11	23.19	23.26	23.34	23.42	23.50
16	23.58	23.66	23.73	23.81	23.89	23.97	24.04	24.12	24.20	24.27
17	24.35	24.43	24.50	24.58	24.65	24.73	24.80	24.88	24.95	25.03
18	25.10	25.18	25.25	25.33	25.40	25.48	25.55	25.62	25.70	25.77
19	25.84	25.92	25.99	26.06	26.13	26.21	26.28	26.35	26.42	26.49
20	26.56	26.64	26.71	26.78	26.85	26.92	26.99	27.06	27.13	27.20
21	27.28	27.35	27.42	27.49	27.56	27.63	27.69	27.76	27.83	27.90
22	27.97	28.04	28.11	28.18	28.25	28.32	28.38	28.45	28.52	28.59
23	28.66	28.73	28.79	28.86	28.93	29.00	29.06	29.13	29.20	29.27
24	29.33	29.40	29.47	29.53	29.60	29.67	29.73	29.80	29.87	29.93
25	30.00	30.07	30.13	30.20	30.26	30.33	30.40	30.46	30.53	30.59
26	30.66	30.72	30.79	30.85	30.92	30.98	31.05	31.11	31.18	31.24
27	31.31	31.37	31.44	31.50	31.56	31.63	31.69	31.76	31.82	31.88
28	31.95	32.01	32.08	32.14	32.20	32.27	32.33	32.39	32.46	32.52
29	32.58	32.65	32.71	32.77	32.83	32.90	32.96	33.02	33.09	33.15

* We are indebted to Dr. C. I. Bliss for permission to reproduce this table, which appeared in *Plant Protection*, No. 12, Leningrad (1937).

(Table A 16 continued on pp. 570-71)

TABLE A 16—(Continued)

%	0	1	2	3	4	5	6	7	8	9
30	33.21	33.27	33.34	33.40	33.46	33.52	33.58	33.65	33.71	33.77
31	33.83	33.89	33.96	34.02	34.08	34.14	34.20	34.27	34.33	34.39
32	34.45	34.51	34.57	34.63	34.70	34.76	34.82	34.88	34.94	35.00
33	35.06	35.12	35.18	35.24	35.30	35.37	35.43	35.49	35.55	35.61
34	35.67	35.73	35.79	35.85	35.91	35.97	36.03	36.09	36.15	36.21
35	36.27	36.33	36.39	36.45	36.51	36.57	36.63	36.69	36.75	36.81
36	36.87	36.93	36.99	37.05	37.11	37.17	37.23	37.29	37.35	37.41
37	37.47	37.52	37.58	37.64	37.70	37.76	37.82	37.88	37.94	38.00
38	38.06	38.12	38.17	38.23	38.29	38.35	38.41	38.47	38.53	38.59
39	38.65	38.70	38.76	38.82	38.88	38.94	39.00	39.06	39.11	39.17
40	39.23	39.29	39.35	39.41	39.47	39.52	39.58	39.64	39.70	39.76
41	39.82	39.87	39.93	39.99	40.05	40.11	40.16	40.22	40.28	40.34
42	40.40	40.46	40.51	40.57	40.63	40.69	40.74	40.80	40.86	40.92
43	40.98	41.03	41.09	41.15	41.21	41.27	41.32	41.38	41.44	41.50
44	41.55	41.61	41.67	41.73	41.78	41.84	41.90	41.96	42.02	42.07
45	42.13	42.19	42.25	42.30	42.36	42.42	42.48	42.53	42.59	42.65
46	42.71	42.76	42.82	42.88	42.94	42.99	43.05	43.11	43.17	43.22
47	43.28	43.34	43.39	43.45	43.51	43.57	43.62	43.68	43.74	43.80
48	43.85	43.91	43.97	44.03	44.08	44.14	44.20	44.25	44.31	44.37
49	44.43	44.48	44.54	44.60	44.66	44.71	44.77	44.83	44.89	44.94
50	45.00	45.06	45.11	45.17	45.23	45.29	45.34	45.40	45.46	45.52
51	45.57	45.63	45.69	45.75	45.80	45.86	45.92	45.97	46.03	46.09
52	46.15	46.20	46.26	46.32	46.38	46.43	46.49	46.55	46.61	46.66
53	46.72	46.78	46.83	46.89	46.95	47.01	47.06	47.12	47.18	47.24
54	47.29	47.35	47.41	47.47	47.52	47.58	47.64	47.70	47.75	47.81
55	47.87	47.93	47.98	48.04	48.10	48.16	48.22	48.27	48.33	48.39
56	48.45	48.50	48.56	48.62	48.68	48.73	48.79	48.85	48.91	48.97
57	49.02	49.08	49.14	49.20	49.26	49.31	49.37	49.43	49.49	49.54
58	49.60	49.66	49.72	49.78	49.84	49.89	49.95	50.01	50.07	50.13
59	50.18	50.24	50.30	50.36	50.42	50.48	50.53	50.59	50.65	50.71
60	50.77	50.83	50.89	50.94	51.00	51.06	51.12	51.18	51.24	51.30
61	51.35	51.41	51.47	51.53	51.59	51.65	51.71	51.77	51.83	51.88
62	51.94	52.00	52.06	52.12	52.18	52.24	52.30	52.36	52.42	52.48
63	52.53	52.59	52.65	52.71	52.77	52.83	52.89	52.95	53.01	53.07
64	53.13	53.19	53.25	53.31	53.37	53.43	53.49	53.55	53.61	53.67
65	53.73	53.79	53.85	53.91	53.97	54.03	54.09	54.15	54.21	54.27
66	54.33	54.39	54.45	54.51	54.57	54.63	54.70	54.76	54.82	54.88
67	54.94	55.00	55.06	55.12	55.18	55.24	55.30	55.37	55.43	55.49
68	55.55	55.61	55.67	55.73	55.80	55.86	55.92	55.98	56.04	56.11
69	56.17	56.23	56.29	56.35	56.42	56.48	56.54	56.60	56.66	56.73

TABLE A 17

ORTHOGONAL POLYNOMIALS

[At the foot of each column the first number is the sum of squares of the polynomial values, the second is the multiplier λ_i needed to give integral values]

$n = 3$		$n = 4$		$n = 5$		$n = 6$		$n = 7$		$n = 8$		$n = 9$		$n = 10$		$n = 11$		$n = 12$	
X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
-1	+1	-3	+1	-2	+2	-5	+5	-5	+5	-9	+9	-20	+20	0	-20	0	-20	0	-20
0	-2	-1	-1	-1	-1	-3	-1	+3	-1	-3	-1	-17	-17	+1	-17	-9	-17	-9	-17
+1	+1	+1	-1	0	-2	0	-2	+1	-1	-3	-1	+2	-8	+2	+7	-7	+7	-7	+7
		+3	+1	+1	+1	+1	+1	+2	+2	+7	+7	+3	+28	+4	+28	+14	+14	+14	+14
2	6	20	4	10	14	70	35/12	10	14	10	10	84	180	70	84	180	28	7/12	252
1	3	2	1	10/3	1	1	5/6	1	1	5/6	1	3/2	5/3	2	3/2	5/3	7/12	21/10	21/10
$n = 7$		$n = 8$		$n = 11$		$n = 12$		$n = 13$		$n = 14$		$n = 15$		$n = 16$		$n = 17$		$n = 18$	
X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
0	-4	0	+6	0	+6	0	+6	0	+6	0	+6	0	+6	0	+6	0	+6	0	+6
+1	-3	-1	+1	+5	-5	+5	-5	+5	-5	+5	-5	+5	-5	+5	-5	+5	-5	+5	-5
+2	0	-1	-7	-4	+1	-4	+1	-4	+1	-4	+1	-4	+1	-4	+1	-4	+1	-4	+1
+3	+5	+1	+3	+1	+7	+1	+7	+1	+7	+1	+7	+1	+7	+1	+7	+1	+7	+1	+7
28	84	6	154	84	168	264	616	168	168	264	616	2184	2772	60	2772	990	2002	468	3/20
1	1	1/6	7/12	7/20	1	2/3	7/12	2	1	2/3	7/12	7/10	3	1	3	5/6	7/12	3/20	3/20
$n = 10$		$n = 11$		$n = 12$		$n = 13$		$n = 14$		$n = 15$		$n = 16$		$n = 17$		$n = 18$		$n = 19$	
X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
+1	-4	-12	+18	+6	-10	0	+6	0	-10	0	+6	0	-10	0	+6	0	-10	0	-10
+3	-3	-31	+3	+11	-9	-14	+4	+4	-9	-14	+4	+4	-9	-14	+4	+4	-9	-14	-31
+5	-1	-35	-17	+1	-6	-23	-1	-1	-6	-23	-1	-1	-6	-23	-1	-1	-6	-23	-35
+7	+2	-14	-22	-14	+2	-1	-6	-6	-1	-22	-6	-6	-1	-22	-6	-6	-1	-22	-14
+9	+6	+42	+18	+6	+6	+30	+6	+3	+6	+30	+6	+3	+6	+30	+6	+3	+6	+30	+42
330	132	8580	2860	780	110	4290	286	156	858	4290	286	156	12012	572	12012	5148	8008	15912	3/20
2	1/2	5/3	5/12	1/10	1	5/6	1/12	1/40	1	5/6	1/12	1/40	3	2	3	2/3	7/24	3/20	3/20

TABLE A 18
TABLE OF SQUARE ROOTS

n	\sqrt{n}	$\sqrt{10n}$	n	\sqrt{n}	$\sqrt{10n}$	n	\sqrt{n}	$\sqrt{10n}$
1.00	1.00	3.16	2.00	1.41	4.47	3.00	1.73	5.48
1.02	1.01	3.19	2.02	1.42	4.49	3.02	1.74	5.50
1.04	1.02	3.22	2.04	1.43	4.52	3.04	1.74	5.51
1.06	1.03	3.26	2.06	1.44	4.54	3.06	1.75	5.53
1.08	1.04	3.29	2.08	1.44	4.56	3.08	1.76	5.55
1.10	1.05	3.32	2.10	1.45	4.58	3.10	1.76	5.57
1.12	1.06	3.35	2.12	1.46	4.60	3.12	1.77	5.59
1.14	1.07	3.38	2.14	1.46	4.63	3.14	1.77	5.60
1.16	1.08	3.41	2.16	1.47	4.65	3.16	1.78	5.62
1.18	1.09	3.44	2.18	1.48	4.67	3.18	1.78	5.64
1.20	1.10	3.46	2.20	1.48	4.69	3.20	1.79	5.66
1.22	1.10	3.49	2.22	1.49	4.71	3.22	1.79	5.67
1.24	1.11	3.52	2.24	1.50	4.73	3.24	1.80	5.69
1.26	1.12	3.55	2.26	1.50	4.75	3.26	1.81	5.71
1.28	1.13	3.58	2.28	1.51	4.77	3.28	1.81	5.73
1.30	1.14	3.61	2.30	1.52	4.80	3.30	1.82	5.74
1.32	1.15	3.63	2.32	1.52	4.82	3.32	1.82	5.76
1.34	1.16	3.66	2.34	1.53	4.84	3.34	1.83	5.78
1.36	1.17	3.69	2.36	1.54	4.86	3.36	1.83	5.80
1.38	1.17	3.71	2.38	1.54	4.88	3.38	1.84	5.81
1.40	1.18	3.74	2.40	1.55	4.90	3.40	1.84	5.83
1.42	1.19	3.77	2.42	1.56	4.92	3.42	1.85	5.85
1.44	1.20	3.79	2.44	1.56	4.94	3.44	1.85	5.87
1.46	1.21	3.82	2.46	1.57	4.96	3.46	1.86	5.88
1.48	1.22	3.85	2.48	1.57	4.98	3.48	1.87	5.90
1.50	1.22	3.87	2.50	1.58	5.00	3.50	1.87	5.92
1.52	1.23	3.90	2.52	1.59	5.02	3.52	1.88	5.93
1.54	1.24	3.92	2.54	1.59	5.04	3.54	1.88	5.95
1.56	1.25	3.95	2.56	1.60	5.06	3.56	1.89	5.97
1.58	1.26	3.97	2.58	1.61	5.08	3.58	1.89	5.98
1.60	1.26	4.00	2.60	1.61	5.10	3.60	1.90	6.00
1.62	1.27	4.02	2.62	1.62	5.12	3.62	1.90	6.02
1.64	1.28	4.05	2.64	1.62	5.14	3.64	1.91	6.03
1.66	1.29	4.07	2.66	1.63	5.16	3.66	1.91	6.05
1.68	1.30	4.10	2.68	1.64	5.18	3.68	1.92	6.07
1.70	1.30	4.12	2.70	1.64	5.20	3.70	1.92	6.08
1.72	1.31	4.15	2.72	1.65	5.22	3.72	1.93	6.10
1.74	1.32	4.17	2.74	1.66	5.23	3.74	1.93	6.12
1.76	1.33	4.20	2.76	1.66	5.25	3.76	1.94	6.13
1.78	1.33	4.22	2.78	1.67	5.27	3.78	1.94	6.15
1.80	1.34	4.24	2.80	1.67	5.29	3.80	1.95	6.16
1.82	1.35	4.27	2.82	1.68	5.31	3.82	1.95	6.18
1.84	1.36	4.29	2.84	1.69	5.33	3.84	1.96	6.20
1.86	1.36	4.31	2.86	1.69	5.35	3.86	1.96	6.21
1.88	1.37	4.34	2.88	1.70	5.37	3.88	1.97	6.23
1.90	1.38	4.36	2.90	1.70	5.39	3.90	1.97	6.25
1.92	1.39	4.38	2.92	1.71	5.40	3.92	1.98	6.26
1.94	1.39	4.40	2.94	1.71	5.42	3.94	1.98	6.28
1.96	1.40	4.43	2.96	1.72	5.44	3.96	1.99	6.29
1.98	1.41	4.45	2.98	1.73	5.46	3.98	1.99	6.31

TABLE OF SQUARE ROOTS—(Continued)

n	\sqrt{n}	$\sqrt{10n}$	n	\sqrt{n}	$\sqrt{10n}$	n	\sqrt{n}	$\sqrt{10n}$
4.00	2.00	6.32	5.00	2.24	7.07	6.00	2.45	7.75
4.02	2.00	6.34	5.02	2.24	7.09	6.02	2.45	7.76
4.04	2.01	6.36	5.04	2.24	7.10	6.04	2.46	7.77
4.06	2.01	6.37	5.06	2.25	7.11	6.06	2.46	7.78
4.08	2.02	6.39	5.08	2.25	7.12	6.08	2.47	7.80
4.10	2.02	6.40	5.10	2.26	7.14	6.10	2.47	7.81
4.12	2.03	6.42	5.12	2.26	7.16	6.12	2.47	7.82
4.14	2.03	6.43	5.14	2.27	7.17	6.14	2.48	7.84
4.16	2.04	6.45	5.16	2.27	7.18	6.16	2.48	7.85
4.18	2.04	6.47	5.18	2.28	7.20	6.18	2.49	7.86
4.20	2.05	6.48	5.20	2.28	7.21	6.20	2.49	7.87
4.22	2.05	6.50	5.22	2.28	7.22	6.22	2.49	7.89
4.24	2.06	6.51	5.24	2.29	7.24	6.24	2.50	7.90
4.26	2.06	6.53	5.26	2.29	7.25	6.26	2.50	7.91
4.28	2.07	6.54	5.28	2.30	7.27	6.28	2.51	7.92
4.30	2.07	6.56	5.30	2.30	7.28	6.30	2.51	7.94
4.32	2.08	6.57	5.32	2.31	7.29	6.32	2.51	7.95
4.34	2.08	6.59	5.34	2.31	7.31	6.34	2.52	7.96
4.36	2.09	6.60	5.36	2.32	7.32	6.36	2.52	7.97
4.38	2.09	6.62	5.38	2.32	7.33	6.38	2.53	7.99
4.40	2.10	6.63	5.40	2.32	7.35	6.40	2.53	8.00
4.42	2.10	6.65	5.42	2.33	7.36	6.42	2.53	8.01
4.44	2.11	6.66	5.44	2.33	7.38	6.44	2.54	8.02
4.46	2.11	6.68	5.46	2.34	7.39	6.46	2.54	8.04
4.48	2.12	6.69	5.48	2.34	7.40	6.48	2.55	8.05
4.50	2.12	6.71	5.50	2.35	7.42	6.50	2.55	8.06
4.52	2.13	6.72	5.52	2.35	7.43	6.52	2.55	8.07
4.54	2.13	6.74	5.54	2.35	7.44	6.54	2.56	8.09
4.56	2.14	6.75	5.56	2.36	7.46	6.56	2.56	8.10
4.58	2.14	6.77	5.58	2.36	7.47	6.58	2.57	8.11
4.60	2.14	6.78	5.60	2.37	7.48	6.60	2.57	8.12
4.62	2.15	6.80	5.62	2.37	7.50	6.62	2.57	8.14
4.64	2.15	6.81	5.64	2.37	7.51	6.64	2.58	8.15
4.66	2.16	6.83	5.66	2.38	7.52	6.66	2.58	8.16
4.68	2.16	6.84	5.68	2.38	7.54	6.68	2.58	8.17
4.70	2.17	6.86	5.70	2.39	7.55	6.70	2.59	8.19
4.72	2.17	6.87	5.72	2.39	7.56	6.72	2.59	8.20
4.74	2.18	6.88	5.74	2.40	7.58	6.74	2.60	8.21
4.76	2.18	6.90	5.76	2.40	7.59	6.76	2.60	8.22
4.78	2.19	6.91	5.78	2.40	7.60	6.78	2.60	8.23
4.80	2.19	6.93	5.80	2.41	7.62	6.80	2.61	8.25
4.82	2.20	6.94	5.82	2.41	7.63	6.82	2.61	8.26
4.84	2.20	6.96	5.84	2.42	7.64	6.84	2.62	8.27
4.86	2.20	6.97	5.86	2.42	7.66	6.86	2.62	8.28
4.88	2.21	6.99	5.88	2.42	7.67	6.88	2.62	8.29
4.90	2.21	7.00	5.90	2.43	7.68	6.90	2.63	8.31
4.92	2.22	7.01	5.92	2.43	7.69	6.92	2.63	8.32
4.94	2.22	7.03	5.94	2.44	7.71	6.94	2.63	8.33
4.96	2.23	7.04	5.96	2.44	7.72	6.96	2.64	8.34
4.98	2.23	7.06	5.98	2.45	7.73	6.98	2.64	8.35

TABLE OF SQUARE ROOTS—(Continued)

n	\sqrt{n}	$\sqrt{10n}$	n	\sqrt{n}	$\sqrt{10n}$	n	\sqrt{n}	$\sqrt{10n}$
7.00	2.65	8.37	8.00	2.83	8.94	9.00	3.00	9.49
7.02	2.65	8.38	8.02	2.83	8.96	9.02	3.00	9.50
7.04	2.65	8.39	8.04	2.84	8.97	9.04	3.01	9.51
7.06	2.66	8.40	8.06	2.84	8.98	9.06	3.01	9.52
7.08	2.66	8.41	8.08	2.84	8.99	9.08	3.01	9.53
7.10	2.66	8.43	8.10	2.85	9.00	9.10	3.02	9.54
7.12	2.67	8.44	8.12	2.85	9.01	9.12	3.02	9.55
7.14	2.67	8.45	8.14	2.85	9.02	9.14	3.02	9.56
7.16	2.68	8.46	8.16	2.86	9.03	9.16	3.03	9.57
7.18	2.68	8.47	8.18	2.86	9.04	9.18	3.03	9.58
7.20	2.68	8.49	8.20	2.86	9.06	9.20	3.03	9.59
7.22	2.69	8.50	8.22	2.87	9.07	9.22	3.04	9.60
7.24	2.69	8.51	8.24	2.87	9.08	9.24	3.04	9.61
7.26	2.69	8.52	8.26	2.87	9.09	9.26	3.04	9.62
7.28	2.70	8.53	8.28	2.88	9.10	9.28	3.05	9.63
7.30	2.70	8.54	8.30	2.88	9.11	9.30	3.05	9.64
7.32	2.71	8.56	8.32	2.88	9.12	9.32	3.05	9.65
7.34	2.71	8.57	8.34	2.89	9.13	9.34	3.06	9.66
7.36	2.71	8.58	8.36	2.89	9.14	9.36	3.06	9.67
7.38	2.72	8.59	8.38	2.89	9.15	9.38	3.06	9.68
7.40	2.72	8.60	8.40	2.90	9.17	9.40	3.07	9.70
7.42	2.72	8.61	8.42	2.90	9.18	9.42	3.07	9.71
7.44	2.73	8.63	8.44	2.91	9.19	9.44	3.07	9.72
7.46	2.73	8.64	8.46	2.91	9.20	9.46	3.08	9.73
7.48	2.73	8.65	8.48	2.91	9.21	9.48	3.08	9.74
7.50	2.74	8.66	8.50	2.92	9.22	9.50	3.08	9.75
7.52	2.74	8.67	8.52	2.92	9.23	9.52	3.09	9.76
7.54	2.75	8.68	8.54	2.92	9.24	9.54	3.09	9.77
7.56	2.75	8.69	8.56	2.93	9.25	9.56	3.09	9.78
7.58	2.75	8.71	8.58	2.93	9.26	9.58	3.10	9.79
7.60	2.76	8.72	8.60	2.93	9.27	9.60	3.10	9.80
7.62	2.76	8.73	8.62	2.94	9.28	9.62	3.10	9.81
7.64	2.76	8.74	8.64	2.94	9.30	9.64	3.10	9.82
7.66	2.77	8.75	8.66	2.94	9.31	9.66	3.11	9.83
7.68	2.77	8.76	8.68	2.95	9.32	9.68	3.11	9.84
7.70	2.77	8.77	8.70	2.95	9.33	9.70	3.11	9.85
7.72	2.78	8.79	8.72	2.95	9.34	9.72	3.12	9.86
7.74	2.78	8.80	8.74	2.96	9.35	9.74	3.12	9.87
7.76	2.79	8.81	8.76	2.96	9.36	9.76	3.12	9.88
7.78	2.79	8.82	8.78	2.96	9.37	9.78	3.13	9.89
7.80	2.79	8.83	8.80	2.97	9.38	9.80	3.13	9.90
7.82	2.80	8.84	8.82	2.97	9.39	9.82	3.13	9.91
7.84	2.80	8.85	8.84	2.97	9.40	9.84	3.14	9.92
7.86	2.80	8.86	8.86	2.98	9.41	9.86	3.14	9.93
7.88	2.81	8.87	8.88	2.98	9.42	9.88	3.14	9.94
7.90	2.81	8.89	8.90	2.98	9.43	9.90	3.15	9.95
7.92	2.81	8.90	8.92	2.99	9.44	9.92	3.15	9.96
7.94	2.82	8.91	8.94	2.99	9.46	9.94	3.15	9.97
7.96	2.82	8.92	8.96	2.99	9.47	9.96	3.16	9.98
7.98	2.82	8.93	8.98	3.00	9.48	9.98	3.16	9.99

A

uthor index

- Abbey, H.—218, 226
 Abelson, R. P.—246, 257
 Acton, F. S.—157, 171
 Anderson, R. L.—380
 Andre, F.—242, 256
 Anscombe, F. J.—322, 332, 338
 Arbous, A. G.—226, 227
 Armitage, P.—247, 248, 257
 Aspin, A. A.—115, 119
 Autrey, K. M.—338

 Baker, P. M.—226
 Balaam, L. N.—274, 298
 Barnard, M. M.—446
 Bartholomew, D. J.—244, 246, 257
 Bartlett, M. S.—296, 297, 298, 328, 338,
 376, 432, 495, 496, 503
 Beadles, J. R.—96, 118
 Beale, H. P.—94, 95, 118
 Becker, E. R.—487, 503
 Beecher, H. T.—446
 Behrens, W. V.—115, 119
 Bennett, B. M.—227
 Berkson, J.—165, 166, 171
 Bernoulli, J.—32
 Best, E. W. R.—226
 Black, C. A.—418
 Bliss, C. I.—327, 569
 Bortkewitch, L. von—225
 Box, G. E. P.—298, 396
 Brandt, A. E.—175, 197, 503
 Breneman, W. R.—102, 118, 152, 171, 503
 Brindley, T. A.—5
 Brooks, S.—532, 539
 Bross, I. D. J.—246, 257, 285, 298
 Brown, B.—503
 Brunson, A. M.—402, 418
 Burnett, L. C.—242, 256
 Burroughs, W.—96, 118
 Butler, R. A.—338

 Caffrey, D. J.—256
 Cannon, C. Y.—338
 Casida, L. E.—134
 Catchpole, H. R.—118
 Chakravarti, I. M.—235, 256
 Chapin, F. S.—118
 Cheeseman, E. A.—256
 Clapham, A. R.—522, 539
 Clarke, G. L.—298, 330, 338
 Cochran, W. G.—90, 115, 118, 119, 226, 255,
 256, 298, 337, 338, 380, 418, 446, 503,
 539
 Collins, E. V.—171
 Collins, G. N.—128, 134
 Corner, G. W.—110, 118
 Cox, D. R.—108, 118
 Crall, J. M.—446
 Crampton, E. W.—118
 Crathorne, A. T.—197
 Crow, E. L.—31
 Culbertson, C. C.—198
 Cushny, A. R.—65

 DasGupta, K. P.—380
 David, F. N.—198
 David, S. T.—198
 Davies, O. L.—380
 Dawes, B.—163, 171
 Dawson, W. T.—457
 Dean, H. L.—118
 Decker, G. C.—242, 256
 Deming, W. E.—517, 539
 DeMoivre, A.—32
 Dixon, W. J.—134
 Doolittle, M. H.—403, 406, 418
 Draper, N. R.—418
 Duncan, D. B.—274, 298, 446
 Duncan, O. D.—418
 Dwyer, P. S.—418
 Dyke, G. V.—501, 503

- Eden, T.—198
 Edwards, A. W. F.—256
 Ehrenkrantz, F.—134
 Eid, M. T.—418
 Evvard, J. M.—175, 198
 Federer, W. T.—338, 492, 503
 Felsenstein, J.—256
 Finney, D. J.—227, 446
 Fisher, C. H.—198
 Fisher, R. A.—60, 65, 90, 108-109, 115, 117, 118, 119, 133, 134, 163, 171, 184, 185, 187, 198, 217, 221, 227, 232, 246, 250, 257, 259, 265, 272, 298, 311-312, 337, 339, 380, 399, 414, 418, 419, 446, 463, 471, 549, 557, 561
 Fitzpatrick, T. B.—539
 Forster, H. C.—337
 Francis, T. J.—226
 Freeman, F. N.—295, 298
 Freeman, H.—31
 Frobisher, M.—226
 Galton, F.—164, 171, 177-178, 198
 Ganguli, M.—291, 298
 Gart, J.—497, 503
 Gates, C. E.—291, 298
 Gauss, C. F.—147, 390, 467, 469
 Geary, R. C.—88, 90
 Goodman, L. A.—502, 503
 Gosset, W. S.—60
 Gowen, J. W.—453, 503
 Gower, J. C.—291, 298
 Gram, M. R.—418
 Graybill, F. A.—65, 134, 418
 Grout, R. A.—188, 198, 418
 Grove, L. C.—96, 118
 Haber, E. S.—198, 377, 379, 380
 Haenszel, W.—256, 257
 Haldane, J. B. S.—241, 256
 Hale, R. W.—338
 Hall, P. R.—487, 503
 Hamaker, H. C.—418
 Hansberry, T. R.—152, 171, 219, 226, 268, 298
 Hansen, M. H.—534, 539
 Harris, J. A.—296, 298
 Harrison, C. M.—134
 Hartley, H. O.—90, 227, 280, 298, 471
 Hasel, A. A.—539
 Healy, M.—338
 Hess, I.—539
 Hiorns, R. W.—470, 471
 Hodges, J. L.—134
 Hoel, P. G.—31
 Holmes, M. C.—247
 Holzinger, K. J.—295, 298
 Hopkins, C. E.—418
 Hotelling, H.—399, 414, 417, 418
 Hsu, P.—227
 Hurwitz, W. N.—534, 539
 Immer, F. R.—529, 539
 Ipsen, J.—246, 257
 Irwin, J. O.—256
 Irwin, M. R.—233, 256
 Iwaskiewicz, K.—119
 James, G. S.—119
 Jessen, R. J.—250, 257, 503
 Kahn, H. A.—503
 Keeping, E. S.—31
 Kempthorne, O.—316, 337, 380, 418, 479, 503
 Kendall, M. G.—134, 194, 195, 198
 Kerrich, J. E.—226, 227
 Keuls, M.—273, 274, 298, 427, 442
 Kimball, A. W.—257
 King, A. J.—539
 Kish, J. F.—471
 Kish, L.—539
 Klotz, J.—134
 Kolodziejczyk, St.—119
 Kurtz, T. E.—275, 298
 Latscha, R.—227
 Lee, A.—171, 172, 175, 196, 197
 Leggatt, C. W.—227, 233, 256
 Lehmann, E. L.—134
 Leslie, P. H.—251, 257
 Leverton, R.—418
 Lewontin, R. C.—256
 Li, H. W.—337
 Lindstrom, E. W.—90, 198, 228, 231, 256
 Link, B. F.—275, 298
 Lipton, S.—380
 Liu, T. N.—337
 Lord, E.—120-122, 128, 134, 553-554
 Lowe, B.—258, 298
 Lush, J. L.—186, 198
 MacArthur, J. W.—27, 31
 McCarty, D. E.—539
 McPeak, M.—539
 Madow, W. G.—539
 Magistad, O. M.—65
 Mahalanobis, P. C.—90, 414, 418
 Mann, H. B.—130, 134, 555
 Mantel, N.—256, 257
 Martin, W. P.—416, 418
 Maxwell, A. E.—418
 May, J. M.—568
 Meier, P.—438, 446

- Meng, C. J.—337
 Merrington, M.—549, 567
 Metzger, W. H.—471
 Mitchell, H. H.—96, 118
 Mitscherlich, A. E.—447, 471
 Molina, E. C.—227
 Monseles, S. P.—337
 Mood, A. M.—65, 134, 418
 Moore, P. G.—122, 134
 Moriguti, S.—284, 298
 Mosteller, F.—226, 328, 338
 Mumford, A. A.—198
 Murphy, D. P.—218, 226

 Newman, D.—273, 274, 298, 427, 442
 Newman, H. H.—295, 298
 Neyman, J.—27, 31, 113, 119

 Ostle, B.—549
 Outhwaite, A. D.—338

 Park, O. W.—118
 Pascal, B.—204, 206, 207
 Patterson, H. D.—468, 471, 501, 503
 Payne, S. L.—539
 Pearl, R.—449, 471
 Pearson, E. S.—65, 90, 227, 280, 298, 471, 552
 Pearson, K.—20, 21, 27, 31, 88, 124, 164, 171, 172, 175, 196, 197, 246, 257
 Pearson, P. B.—118
 Peebles, A. R.—65
 Penquite, R.—471
 Pesek, I.—418
 Pillai, K. C. S.—120, 134
 Pitman, E. J. G.—196, 197, 198
 Poisson, S. D.—223, 226
 Porter, R. H.—337, 380
 Price, W. C.—453

 Rao, C. R.—235, 256, 418
 Reed, J. F.—539
 Reëd, L. J.—449, 471
 Richardson, C. H.—152, 171, 218–219, 226, 268, 298
 Richardson, R.—298
 Riedel, D. C.—539
 Rigney, J. A.—539
 Roberts, H.—134, 418
 Rogers, M.—253
 Rourke, R. E. K.—226
 Rutherford, A.—338

 Salisbury, G. W.—290
 Sampford, M. R.—539
 Satterthwaite, F. E.—338, 380
 Saunders, A. R.—380

 Scattergood, L. W.—539
 Scheffé, H.—271, 298, 338
 Schlottfeldt, C. S.—338
 Serfling, R. E.—539
 Sheppard, W. F.—83, 90
 Sherman, I. L.—539
 Shine, C.—291, 298
 Silver, G. A.—503
 Sinha, P.—380
 Slonim, M. J.—539
 Smirnov, N. V.—90
 Smith, A. H.—118, 459, 471
 Smith, C. A. B.—418
 Smith, C. E.—256
 Smith, G. M.—412, 446
 Smith, H.—418
 Smith, H. F.—428, 446
 Smit., S. N.—104, 118
 Snedecor, G. W.—31, 152, 171, 198, 240, 256, 265, 298, 379, 380, 503
 Snedecor, J. G.—28
 Snell, M. G.—198
 Spearman, C.—194, 198
 Sprague, G. F.—446
 Stephan, F. F.—539
 Stevens, W. L.—468, 470, 471, 492, 503
 Stewart, R. T.—171
 Strand, N. V.—250, 257, 503
 Stuart, A.—134, 198, 539
 Swanson, P. P.—118, 171, 418, 446, 459, 471

 Talley, P. J.—65
 Tam, R. K.—65
 Tang, P. C.—280, 298
 Theriault, E. J.—471
 Thomas, G. B., Jr.—226
 Thompson, C. M.—551, 567
 Tippet, L. H. C.—65
 Trickett, W. H.—119
 Tukey, J. W.—246, 257, 275, 298, 322, 331–334, 337, 338

 Vasey, A. J.—337
 Vos, B. J.—457

 Wald, A.—290, 298
 Walker, C. B.—226
 Walker, R. H.—118
 Wallace, D. L.—275, 298
 Walser, M.—446
 Welch, B. L.—115, 119
 Wentz, J. B.—171
 West, Q. M.—518, 539
 Westmacott, M.—338
 White, C.—130, 131, 134, 556
 Whitney, D. R.—130, 134, 555
 Wiebe, G. A.—84, 90

580 *Author Index*

- Wilcoxon, F —128–130, 134 555, 556
Wilk, M B —479, 503
Williams, C B.—330, 338
Williams, E J.—399, 418
Williams, R E O —247
Willier, J G —402, 418
Wilsie, C P —380
Winsor, C P —298, 330, 338
Woolsey, T D —539
Wright, E B —131, 134
Wright, S —418
Yates, F —119, 247, 257, 265, 298, 337, 338,
342, 380, 446, 471, 488, 501, 503, 539
Youden, W J —94, 95, 118
Young, M —198
Youtz, C —328, 338
Yule, G U —189, 198
Zarcovich, S S —539
Zelen, M —492, 503
Zoellner, J A —418
Zweifel, J R —497, 503

Index to numerical examples analyzed in text

(The index is arranged by the statistical technique involved. The type of data being analyzed is described in parentheses.)

Additivity, Tukey's test

Latin squares (responses of monkeys to stimuli), 335

two-way classification (numbers of insects caught in light trap), 333

Analysis of covariance

in one-way classification, computations

one X -variable (leprosy patients, scores for numbers of bacilli), 422

two X -variables (rate of gain, age, and weight of pigs), 440

in two-way classification, computations

one X -variable (mental activity scores of students), 426

(yields and plant numbers of corn), 428

two X -variables (yields, heights, and plant numbers of wheat), 444

interpretation of adjustments (per capita incomes and expenditures per pupil in schools)
431

Asymptotic regression, fitting (temperatures in refrigerated hold), 469

Binomial distribution

fitting to data (random digits), 205

see also Proportions, analysis of

Bivariate normal distribution, illustration (heights and lengths of forearm of men), 177

Cluster sampling, estimation of proportions (numbers of diseased plants), 514

Components of variance

nested classification, estimation of components

equal sizes (calcium contents of turnip greens), 286

unequal sizes (wheat yields of farms), 292

one-way classification, estimation of components

equal sizes (calcium contents of turnip greens), 281

unequal sizes (percent of conceptions to inseminations in cows), 290

Correlation

comparison and combination of r 's and r - z transformation (initial weights and gains in weight of steers), 187

computation and test of r (heights of brothers and sisters), 172

intraclass computations (numbers of ridges on fingers of twins), 295

partial computations (age, blood pressure and cholesterol level of women), 401

rank correlation coefficient, computations (rated condition of rats), 194

Discriminant function, computations (presence or absence of *Azotobacter* in soils), 416

582 *Index to Numerical Examples Analyzed in Text*

Exponential growth curve, fitting (weights and ages of chickens), 450

Factorial experiments, analysis

2×2 , interaction absent (riboflavin concentration of collard leaves), 343

2×2 , interaction present (gains in weight of pigs), 346

3×2 , (gains in weight of rats), 347

$2 \times 2 \times 2$ and $2 \times 3 \times 4$ (gains in weight of pigs), 359, 362

Kurtosis, test of (numbers of inhabitants of U.S. cities), 87

Latin square

 analysis (yields of millet for different spacings), 313

 missing value, estimation (milk yields of cows), 272

Least significant difference (LSD) (doughnuts), 272

Mean

 computation from frequency distribution (weights of swine), 82

 estimation and confidence interval (vitamin C content of tomato juice), 39

Median, estimation and confidence interval (days from calving to oestrus in cows), 123

Missing values, estimation and analysis

 Latin square (milk yields of cows), 319

 two-way classification (yields of wheat), 318

Nested (split-plot) design, analysis (yields of alfalfa), 371

Nested classifications, analysis for mixed effects model (gains in weight of pigs), 289. *See also* Components of variance.

Newman-Keuls test (grams of fat absorbed by doughnuts), 273

Normal distribution

 confidence interval for mean (σ unknown) (vitamin C content of tomato juice), 39

 tests of skewness and kurtosis (numbers of inhabitants of U.S. cities), 85–87

One-way classification, frequencies

 examination of variation between and within classes (numbers of insect larvae on cabbage), 234

 test of equality of frequencies (random digits), 232

 test of estimated frequencies (numbers of weed seeds in meadow grass), 237

 test of specified frequencies (Mendelian) (color of crosses of maize), 228

One-way classification, measurements *See also* Components of variance.

 analysis of variance

 more than two classes (grams of fat absorbed by doughnuts), 259

 samples of unequal sizes (survival times of mice with typhoid), 278

 two classes (comb weights of chickens), 267

 standard error of comparison among class means (yields of sugar), 269

Ordered classifications, analysis by assigned scores (health status and degree of infiltration of leprosy patients), 245

Orthogonal polynomials, fitting (weights of chick embryos), 461

Paired samples, comparison of means

 measurements (lesions on tobacco leaves), 95

 proportions (diphtheria bacilli on throats of patients), 213

Partitioning of Treatments sums of squares

 (area/weight ratio of leaves of citrus trees), 309

 by orthogonal polynomials (yields of sugar), 350

 in factorial experiment (gains in weight of rats), 349

 (soybean seeds, failures to germinate), 308

Perennial experiment, analysis (weights of asparagus), 378

Poisson distribution

- fitting (weed seeds in meadow grass), 224
- homogeneity tests (deaths of chinch bugs under exposure to cold), 242
- test of goodness of fit (weed seeds), 237
- variance test (random digits), 232

Proportions, analysis of

- confidence interval (fields sprayed for corn borer), 5
- in one-way classification, *see* Two-way classification, frequencies
- in two-way classification
 - 2×2 table (percent survival of plum root-stocks), 495
 - 2×3 table (percent of children and parents with emotional problems), 497
 - $R \times C$ table (in logs) (death rates of men by age and numbers of cigarettes smoked), 498

Range

- analog of t -test (numbers of worms in rats), 121
- estimation of σ from (vitamin C content of tomato juice), 39

Ranks

- signed rank test (Wilcoxon) (lengths of corn seedlings), 129
- two-sample sum of ranks test (Mann-Whitney)
 - equal sizes (numbers of borer eggs on corn plants), 130
 - unequal sizes (survival times of cats and rabbits), 131

Rank correlation coefficient (rated condition of rats), 194

Ratios, estimation (sizes and corn acres in farms), 168

Regression

- comparison of "between classes" and "within classes" regressions (scores for bacilli in leprosy patients), 437
- comparison of regression in two samples (age and cholesterol concentration of women)
433
- fitted to treatment means (yields of millet), 314
- fitting of linear
 - (age and blood pressure of women), 136
 - (percent wormy fruits and size of crop of apple trees), 150
- fitting of quadratic (protein content and yield of wheat), 454
- multiple, fitting for 2 and 3 X -variates (phosphorus contents of soils), 384, 405
- test for linear trend in proportions (leprosy patients), 247
- test of intercept (speed and draft of ploughs), 167
- test of linearity (survival time of cats with ouabain), 458

Rejection of observations, application of rule (yields of wheat), 318

Response curves, two-factor experiments (yields of cowpea hay), 352

Response surface, fitting (ascorbic acid content of snapbeans), 354

Sample size estimation (yields of wheat), 417

- in two-stage sampling (percent of sugar in sugar-beets), 517

Series of experiments, analysis (numbers of soybean plants), 377

Sets of 2×2 tables, analysis (problem children in school and previous infant losses (mothers), 253

Sign test (ranking of beef patties), 126

Skewness, test of (numbers of inhabitants in U.S. cities), 85

Split-plot experiment, analysis (yields of alfalfa), 371

Standard deviation, computation (vitamin C content of tomato juice), 39

- from frequency distribution (weights of swine), 82

Stratified random sampling

- optimum allocation (numbers of students in colleges), 524
- standard error of mean (wheat yields), 522
- with attributes and proportions (numbers of vegetable gardens), 527

- variance test of homogeneity, 240–242
- Bivariate normal distribution, 177–179
- Blocks, 299
 - efficiency of blocking, 311
- Case study, 152
- Central limit theorem, 51, 209
- Chi-square (χ^2), 20–26, 30, 212
 - correction for continuity, 125, 209–210
 - distribution of, 22–26, 73
 - in goodness of fit tests, 236–238
 - in $R \times C$ contingency tables, 250–253
 - in tests of Mendelian ratios, 228–231, 248–250
 - in $2 \times C$ contingency tables, 238–240
 - in 2×2 contingency tables, 215–220
 - in variance test for binomial, 240–243
 - in variance test for Poisson, 231–233
 - normal approximation to, 233
 - relation to distribution of sample variance s^2 , 73–74
 - table of, 550–551
 - test of binomial proportion, 20–22, 213–214
- Class
 - interval, 23
 - mark, 67, 73, 82
- Cluster sampling, 511
 - formulas in simple cluster sampling, 513–515
- Coding, 81
- Coefficient of variation, 62
- Common elements, 181
- Comparison
 - among more than two means, 268–275
 - definition, 269
 - of all pairs of means, 271–275
 - of mean scores, 244–245
 - of observed and expected frequencies
 - more than two classes, 228–238
 - two classes, 20–27
 - of two means in independent samples, 100–105, 114–116
 - of two means in paired samples, 93–95, 97–99
 - of two proportions in independent samples, 215–223
 - of two proportions in paired samples, 213–215
 - orthogonal, 309
 - rule for standard error, 269, 301–302
- Components of variance, 280
 - in factorial experiments, 364–369
 - in three-stage sampling, 285–288, 291–294
 - in two-stage sampling, 280–285, 289–291, 529–533
 - confidence limits, 284–285
- Compound interest law, 447
- Confidence intervals, 5–7, 14–15, 29
 - for an individual Y , given X , 155–157
 - for binomial proportion, 210–211
 - for components of variance, 284–285
 - for correlation coefficient, 185–188
 - for partial regression coefficients, 391
 - for population mean (σ known), 56
 - for population mean (σ unknown), 61, 122
 - for population median, 124–125
 - for population regression line, 153–155
 - for population variance, 74–76
 - for ratio of two variances, 197
 - for slope in regression, 153
 - one-sided, or one-tailed, 57
 - table for binomial distribution, 6–7
 - upper and lower, 58
- Confidence limits, 5–7. *See also* Confidence intervals.
 - upper and lower, 58
- Contingency table
 - $R \times C$, 250–252
 - $2 \times C$, 238–243
 - 2×2 , 215–223
 - sets of 2×2 tables, 253–256
- Continuity correction, 125, 209–210, 230–231
- Continuous distribution, 23
- Correction
 - for continuity, 125, 209–210
 - for finite size of population, 513
 - for mean, 261–262
 - for working mean, 47–48
 - Sheppard's, 83
- Correlation
 - and common elements, 181–183
 - calculation in large sample, 190–193
 - coefficient, 172
 - combination of separate estimates, 187
 - comparison of several coefficients, 186
 - confidence interval for, 185
 - tables, 557–559
 - tests of significance, 184–188
 - intraclass, 294
 - multiple, 402
 - nonsense, 189
 - partial, 400–401
 - rank, 193–195
 - relation to bivariate normal distribution, 177–179
 - relation to regression, 175–177
 - role in selection, 189
 - role in stream extension, 189
 - utility of, 188–190
- Covariance, 181. *See also* Analysis of covariance.
- Curve fitting, 447–471

- Degrees of freedom, 45
 - for chi-square
 - in contingency tables, 217, 239, 251
 - in goodness of fit tests, 237
 - in tests of homogeneity of variance, 297
 - in analysis of variance
 - Latin square, 314
 - one-way classification, 261
 - two-way classification, 301, 307
 - in correlation, 184
 - in regression, 138, 145, 162–163, 385
- Deletion of a variable, 412
- Dependent variable in regression, 135
- Design of investigations
 - comparison of paired and independent samples, 106–109
 - efficiency of blocking, 311–312
 - factorial experiments, 339–364
 - independent samples, 91, 100–106, 114–116, 258–275
 - Latin squares, 312–317
 - Missing data, 317–321
 - paired samples, 91–99
 - perennial crops, 377–379
 - randomized blocks or groups, 299–310
 - role of randomization, 109–111
 - sample size, 111–114, 221–223
 - sample surveys, 504
 - series of experiments, 375–377
 - two-stage (split-plot or nested) designs, 369–375
 - use of covariance, 419–432
 - use of regression, 135
- Deviations
 - from sample mean, 42
- Digits
 - random, 12
 - table of, 543–546
- Discrete distribution, 16
- Discriminant function, 414
 - computations, 416–418
 - relation to multiple regression, 416
 - uses, 414
- Distance between populations, 415
- Distribution. *See also* the specific distribution.
 - binomial, 17
 - bivariate normal, 177
 - chi-square, 73
 - F (variance ratio), 117
 - multinomial, 235
 - normal, 32
 - Poisson, 223
 - Student's t -, 59
- Dummy variable, 416
- analysis of covariance, 423–424, 427
- Latin squares, 316
- randomized blocks, 311
- range, 46
- rank tests, 132
- sign test, 127
- Equally likely outcomes, 199
- Error
 - of first kind (Type I), 27, 31
 - of measurement
 - effect on estimates in regression, 164–166
 - of second kind, (Type II), 27, 31
 - regression, 421
 - standard (*See* Standard error.)
- Estimate or estimator
 - interval, 5, 29
 - point, 5, 29
 - unbiased, 45, 506
- Expected numbers, 20, 216, 228–240
 - minimum size for χ^2 tests, 235, 241
- Experiment. *See* Design of investigations.
- Experimental sampling, used to illustrate
 - binomial confidence limits, 14
 - binomial frequency distribution, 16
 - central limit theorem, 51–55
 - chi-square (1 d.f.) for binomial, 22–26
 - confidence interval for population mean μ , 78–79
 - distribution of sample means from a normal distribution, 70–72
 - distribution of sample standard deviation s , 72–73
 - distribution of sample variance s^2 , 72–73
 - F -distribution, 266
 - t -distribution, 77–78
- Exponential
 - decay curve, 447
 - growth curve, 447, 449–453
- Extrapolation, 144, 456
- F -distribution, 117
 - effect of correlated errors, 323
 - effect of heterogeneous errors, 324
 - effect of non-normality, 325
 - one-tailed tables, 560–567
 - two-tailed table, 117
- Factor, 339
- Factorial experiment, 339
 - analysis of 2^2 factorial
 - interaction absent, 342–344
 - interaction present, 344–346
 - analysis of 2^3 factorial, 359–361
 - analysis of general three-factor experiment, 361–364
 - analysis of general two-factor experiment, 346–349

- compared with single-factor experiment, 339–342
 - fitting of response curves to treatments, 349–354
 - fitting of response surface, 354–358
- Finite population correction, 513
- First-order reaction curve, 448. *See also* Asymptotic regression.
- Fixed effects model
 - in factorial experiments, 364–369
 - in one-way classification, 275
- Fourfold (2×2) table, 215
- Freedom, degrees of. *See* Degrees of freedom.
- Frequency
 - class, 23
 - cumulative, 26
 - distribution, 16, 30
 - continuous, 23
 - discrete, 16
 - number of classes needed, 80–81
 - expected, 20
 - observed, 20
- g_1 and g_2 tests for non-normality, 86–87
- Genetic ratios
 - tests of, 228–231, 248–249
- Geometric mean, 330
- Goodness of fit test, χ^2 , 84. *See also* Chi-square.
- Graphical representation, 16, 40
- Grouping
 - loss of accuracy due to, 81
- Growth curve
 - exponential, 449
 - logistic, 448–449
- Harmonic mean, 475
- Heterogeneity
 - chi-square, 248
 - of variances, 296, 324
- Hierarchical classifications, 285–289
- Histogram, 25
- Homogeneity, test of
 - in binomial proportions, 240
 - in Poisson counts, 231
 - in regression coefficients, 432
 - of between- and within-class regressions, 436
- Hotelling's T^2 -test, 414, 417
- Hypotheses about populations, 20. *See* Tests of significance.
 - null, 26, 30
 - tests of
 - in analysis of variance, 323
 - in binomial distribution, 201
 - in probability, 201
 - with attributes, 219
- Independent samples
 - comparison of two means, 100–105, 114–116
 - comparison of two proportions, 215–223
- Independent variable in regression, 135
- Inferences about population, 3–9, 29, 504–505. *See also* Confidence intervals.
- Interaction, 341
 - possible reasons for, 346
 - three-factor, 359–364
 - in contingency tables, 496
 - two-factor, 341–349, 473
- Interpolation in tables, 541
- Interval estimate, 5, 29. *See also* Confidence interval.
- Intraclass correlation, 294–296
- Inverse matrix, 389, 403, 409–412
- Kendall's τ , 194
- Kurtosis, 86
 - effect on variance of s^2 , 89
 - test for, 86–88
 - table, 552
- Latin square, 312
 - efficiency, 316
 - model and analysis of variance, 312–315
 - rejection of observations, 321–323
 - test of additivity, 334–337
- Least significant difference; 272
- Least squares, method of, 147
 - as applied to regression, 147
 - Gauss theorem, 147
 - in two-way tables with unequal numbers, 483–493
- Level of significance, 27
- Limits, confidence. *See* Confidence intervals.
- Likelihood, maximum, 495
- Linear calibration, 159–160
- Linear regression. *See* Regression
- Listing, 509–511
- Logarithm
 - common and natural, 451–452
- Logarithmic
 - graph paper, 450, 452
 - transformation, 329–330
- Logistic growth law, 448–449
- Logit transformation, 494, 497–503
- Lognormal distribution, 276
- Main effect, 340–342
- Main plot, 369
- Mann-Whitney test, 130
 - significance levels, 131, 555–556

- Mantel-Haenszel test, 255–256
- Mathematical model for
 analysis of covariance, 419
 exponential growth curve, 449
 factorial experiment, 357, 364–369
 Latin square, 313
 logistic growth curve, 448–449
 multiple regression, 382, 394
 nested (split-plot) designs, 370
 one-way classification
 fixed effects, 275
 mixed effects, 288
 random effects, 279, 289
 orthogonal polynomials, 460–465
 regression, 141
 asymptotic, 468
 non-linear, 465
 two-way classification, 302–308, 473
- Matrix, 390
 inverse, 390, 409, 439, 490
- Maximin method, 246
- Maximum likelihood, 495
- Mean
 absolute deviation, 44
 adjusted, 421, 429
 arithmetic, 39
 correction for, 261–262
 distribution of, 51
 geometric, 330
 harmonic, 475
 weighted, 186, 438, 521
- Mean square, 44
 expected value
 in factorial experiments, 364–369
 with proportional sub-class numbers, 481–482
- Mean square error
 in sampling finite populations, 506
- Measurement data, 29
- Median 123
 calculation from large sample, 123
 confidence interval, 124–125
 distribution of sample median, 124
- Mendelian inheritance
 heterogeneity χ^2 test, 248–249
 test of specified frequencies, 228–231
- Missing data
 in Latin square, 319–320
 in one-way classification, 317
 in two-way classification, 317–321
- Mitscherlich's law, 447. *See also* Asymptotic regression.
- Mixed effects model
 in factorial experiments, 364–369
 in nested classifications, 288–289
- Mode, 124
- Model. *See* Mathematical model.
- Model I, fixed effects. *See* Fixed effects model.
- Model II, random effects. *See* Random effects model.
- Moment about mean, 86
- Monte Carlo method, 13
- Multinomial distribution, 235
- Multiple comparisons, 271–275.
- Multiple covariance. *See* Analysis of covariance.
- Multiple regression. *See* Regression.
- Multiplication rule of probability, 201
- Multivariate *t*-test, 414, 417
- Mutually exclusive outcomes, 200
- Nested
 classifications, 285–289, 291–294
 designs, 369
- Newman-Keuls test, 273–275
- Non-additivity
 effects of in analysis of variance, 330–331
 removal by transformation, 329, 331
 tests for
 in Latin square, 334–337
 in two-way classification, 331–334
- Non-parametric methods
 Mann-Whitney test, 130
 median and percentiles, 123–125
 rank correlation, 193–195
 sign test, 127
 Wilcoxon signed rank test, 128
- Normal distribution, 32
 formula for ordinate, 34
 mean, 32
 method of fitting to observed data, 70–72
 reasons for use of, 35
 relation to binomial, 32, 209–213
 standard deviation, 32
 table of cumulative distribution, 548
 table of ordinates, 547
 tests of normality, 86–88
- Normal equations, 383
 in multiple regression, 383, 389, 403
 in two-way classifications, 488–491
- Normality, test of, 84–88
- Null hypothesis, 26, 30
- One-tailed tests, 76–77, 98–99
- One-way classification, frequencies
 expectations equal, 231–235, 242–243
 expectations estimated, 236–237
 expectations known, 228–231
 expectations small, 235
- One-way classification, measurements
 analysis of variance, 238–248

- comparisons among means, 268–275
- effects of errors in assumptions, 276–277
- model I, fixed effects, 275
- model II, random effects, 279–285, 289–291
- rejection of observations, 321–323
- samples of unequal sizes, 277–278
- Optimum allocation
 - in stratified sampling, 523–526
 - in three-stage sampling, 533
 - in two-stage sampling, 531–533
- Ordered classifications
 - methods of analysis, 243–246
- Order statistics, 123
- Orthogonal comparisons, 309
 - in analysis of factorial experiments, 346–361
- Orthogonal polynomials, 349–351, 460–464
 - tables of coefficients (values), 351, 572
- Outliers (suspiciously large deviations)
 - in analysis of variance, 321
 - in regression, 157
- Paired samples, 91
 - comparison of means, 93–95, 97–99
 - comparison of proportions, 213–215
 - conditions suitable for pairing, 97
 - self-pairing, 91
 - versus independent samples, 106–108
- Parabolic regression, 453–456
- Parameter, 32
- Partial
 - correlation, 400
 - coefficient, 400
 - regression coefficient, 382
 - interpretation of, 393–397
 - standard, 398
- Pascal's triangle, 204
- Percentages, analysis of. *See* Proportions, analysis of.
- Percentiles, 125
 - estimation by order statistics, 125
- Perennial experiments, 377–379
- Placebo, 425
- Planned comparisons, 268–270
- Point estimate, 5, 29
- Poisson distribution, 223–226
 - fitting to data, 224–225
 - formula for, 223
 - test of goodness of fit, 236–237
 - variance test of homogeneity, 232–236
- Polynomial regression or response curve, 349–354
- Pooling (combining)
 - correlation coefficients, 187
 - estimated differences in 2×2 tables, 254–256
- estimates of variance, 101–103
- of classes for χ^2 tests, 235
- regression coefficients, 438
- Population, 4, 29, 504–505
 - finite, 504–505, 512–513
 - sampled, 15, 30
 - target, 30
- Power function, 280
- Primary sampling units, 528
- Probability
 - simple rules, 199–202, 219
- Probability sampling, 508–509
- Proportional sub-class numbers, method of, 478–483
- Proportions, analysis of
 - in one-way classifications, 240–243
 - test for a linear trend, 246–248
 - in two-way classifications, 493
 - in angular (arcsin) scale, 496
 - in logit scale, 497–503
 - in original (p) scale, 495–497
 - in sets of 2×2 tables, 253–256
- Random digits (numbers), 12–13, 30
 - table, 543–546
- Random effects model
 - in factorial experiments, 364–369
 - in one-way classification, 279–294
- Randomization, 110
 - as precaution against bias, 109–111
- Randomization test (Fisher's), 133
- Randomized blocks, 299. *See also* Two-way classifications.
- efficiency of blocking, 311
- Random sampling, 10–11, 30
 - stratified, 11
 - with replacement, 11
 - without replacement, 11, 505
- Range, 39
 - efficiency relative to standard deviation, 46
 - relation to standard deviation, 40
- Studentized Range test, 272–273
- t -test based on, 120
 - tables, 553–554
 - use in comparison of means, 275
- Rank correlation, 193–195
- Ranks, 128
 - efficiency relative to normal tests, 132
 - rank sum test, 130–132
 - signed rank test, 128–130
- Ratio
 - estimates in sample surveys, 536–537
 - estimation of, 170
 - standard error of, 241, 515, 537
- Rectangular (uniform) distribution 81

Rectification, 449

Regression, 135

analysis of variance for, 160–163

coefficient (slope), 136

interval estimate of, 153

value in some simple cases, 147–148

comparison of “between classes” and
“within classes” regressions, 436–438

comparison of regression lines, 432–436

confidence interval for slope, 153

deviations from, 138

effects of errors in X , 164–166

estimated regression line, 144–145

estimated residual variance, 145–146

estimates in sample surveys, 537–538

equation, 136

historical origin of the term, 164

in one-way classification of frequencies,
234

line through origin, 166–169

linear regression of proportions, 246–248

mathematical model, 141–144

multiple, 381

computations in fitting, 383–393, 403–
412

deletion of an independent variable, 412

deviations mean square, 385–389

effects of omitted variables, 394–397

importance of different X -variables,
398–400

interpretation of coefficients, 393–397

partial regression coefficient, 382

prediction of individual observation,
392

prediction of population line, 392

purposes, 381

selection of variates for prediction, 412–
414

standard error of a deviation, 392

standard errors of regression coeffi-
cients, 391

testing a deviation, 392–393

tests of regression coefficients, 386–388

non-linear in some parameters, 465–471

general method of fitting, 465–467

parabolic, 453–456

prediction of individual observation, 155–
157

prediction of the population line, 153

prediction of X from Y , 159–160

relation to correlation, 175–177, 188–190

shortcut computation, 139

situation when X varies from sample to
sample, 149–150

testing a deviation, 157–158

tests for linearity, 453–459

Rejection of observations

in analysis of variance, 321–323

Relative amount of information, 311

Relative efficiency, 46

of range, 46

Relative rate of increase, 450

Replications, 299

Residuals, 300–301, 305–307

Response curve

polynomial, 349–351

Response surface, 346

example of fitting, 354–358

Ridits, 246

Rounding errors, 81

effect on accuracy of \bar{X} and s , 81

Sample, 4, 29

cluster, 511, 513–515

non-random, 509

probability, 508–509

random, 10–11, 30, 505, 511

stratified random, 507, 520–527

systematic, 519

Sample mean, \bar{X} , 39

calculation from a frequency distribution,
80–83

frequency distribution of, 51

Sample standard deviation s , 44

Sampling fraction, 512

unequal, 507

Sampling unit, 509

Scales with limited values, 132

Scheffe's test, 271

Scores

assigned to ordered classifications, 244–
246

Selection of candidates, 189

Selection of variates for prediction, 412–414

Self-pairing, 91, 97

Self-weighting estimate, 521

Semi-logarithmic graph paper, 450

Series of experiments, 375–377

Sets of 2×2 tables, 253–256

Sheppard's correction, 83

Sign test, 125–127

efficiency of, 127

table of significance levels, 554

Signed rank test, 128

significance levels, 129, 555

Significance

level, 27

tests of (*See Tests of significance.*)

Simple random sampling, 505–507

of cluster units, 513–515

properties of estimates, 511–515

size of sample, 516–518

Size of sample

for comparing two proportions, 221–222

- for estimating population mean, 58
- for tests of significance when comparing means 111–114
- in sampling finite populations, 516–518
- in two-stage (nested) sampling, 281
- within strata, 523–526
- Skewness, 72
 - test of, 86
 - table, 252
- Smoothing, 447
- Spearman's rank correlation coefficient, 194
- Split-plot (nested) design, 369
 - analysis of variance, 370–373
 - comparison with randomized blocks, 373
 - reasons for use, 369–370
- Square roots
 - method of finding, 541
 - table, 573–575
- Square root transformation, 325–327
- Standard deviation
 - of estimates from data
 - adjusted difference, 423
 - difference, 100, 104, 106, 115, 190
 - g_1 for skewness, 86
 - g_2 for kurtosis, 87
 - mean of random sample, 50, 512
 - median, 124
 - population total, 51, 513
 - regression coefficient, 138, 391
 - sample total, 51
 - sum, 190
 - transformed correlation, 185
 - variance, 89
 - of population
 - binomial, 207–208
 - normal, 32
 - Poisson, 225
- Standard error, 50 *See also* Standard deviation
- Standard normal deviate, 36
- Standard normal variate 36
- Standard partial regression coefficient, 398
- Step up and step down methods, 413
- Stratified random sampling, 11, 507, 520
 - for attributes 526–527
 - optimum allocation, 523–526
 - proportional allocation, 521–523
 - reasons for use, 520
- Stream extension, 189
- Structural regression coefficient, 165
- Studentized Range test 272–273
 - shortcut computation using ranges, 275
 - table 568
- Student's t -distribution 59
 - table 549
- Sub-class numbers
 - equal 475
- equal within rows 477
- proportional, 478
- unequal, 472
- Sub-plots, 369
- Sub-sampling *See* Two-stage sampling
- Sum
 - of products, 136
 - correction for means, 141
 - of squares, 44–45
 - correction for mean, 48–49
- Systematic sampling, 519
- t (Student's t -distribution) 59
 - table, 549
- Tests of significance, 26–30
 - goodness of fit test, χ^2 , 84–85
 - in analysis of covariance, 423–425
 - in $R \times C$ contingency tables 250–252
 - in $2 \times C$ contingency tables 238–243
 - 246–249
 - binomial proportion, 26–28, 211–213
 - all differences among means 271–275
 - correlation coefficient, 184–188
 - difference between means of independent samples, 100–105 114–116
 - difference between means of paired samples, 93–95, 97–99
 - difference between two binomial proportions, 213–221
 - equality of two correlated variances 195–197
 - equality of two variances, 116
 - goodness of fit of distributions, 236–237
 - homogeneity of Poisson samples, 232–236
 - homogeneity of variances 296–298
 - linear trend in proportions, 246–248
 - linearity of regression, 453–460
 - multiple correlation coefficient 402
 - rank correlation coefficient 194
 - single classification with estimated frequencies, 236–238
 - single classification with equal frequencies 231–234
 - single classification with specified frequencies, 228–231
 - t -test based on range 120
 - test of skewness, 86
 - tests of kurtosis, 86–88
- Three-stage sampling, 285–288, 533
 - allocation of sample sizes 533
- Transformation, 277
 - logarithmic, 329–330
 - logit 494–497–503
 - to remove non-additivity 331–332
 - to stabilize variance 325
 - angular (arcsin) 327–329
 - square root 325–327

- use in fitting non-linear relations 448-453
- Treatments, definition, 91
- Treatment combination, 340
- Tukey's tests for additivity 331-337
- Two-stage sampling 528
 - reasons for use 528
 - with primary units of equal size 529-533
 - choice of sample and sub-sample sizes, 531-533
 - with primary units of unequal sizes 534-536
- Two-way classifications, frequencies, 238-243
 - $R \times C$ tables 250-253
 - sets of 2×2 tables 253 257
 - $2 \times C$ tables 238 243 246 250
 - 2×2 (fourfold) tables 215 223
- Two-way classifications measurements
 - additivity assumption, 302, 330-334
 - analysis of variance 299-301
 - mathematical model 302 307
 - rejection of observations 321 323
 - test of additivity 331 334
 - with unequal numbers 472
 - complications involved 472 475
 - equal weights within rows 477 478
 - least squares analysis $R \times C$ table 488-493
 - method of proportional numbers, 478-483
 - $R \times 2$ table, 484-487
 - 2×2 table, 483 484
 - unweighted analysis, 475-477
- Two-way classifications proportions
 - analysis in logit scale 497 503
 - analysis in proportions scale 495-497
 - approaches to analysis, 493-495
- Unbiased estimate, 45
- Uniform distribution 51
 - in relation to rounding errors, 81
- Unweighted means, method of, 475-477
- Variance, 53
 - analysis (See Analysis of variance)
 - comparison of two correlated variances, 195-197
 - comparison of two variances, 116
 - components (See Components of variance)
 - confidence interval for, 74
 - of difference, 100, 104, 106, 115, 190
 - of sum, 190
 - ratio F , 265
 - distribution under general hypothesis 280
 - table 560-567
 - test of homogeneity 296-298
- Variation coefficient of 62
- Weighted mean
 - in stratified sampling 521
 - of differences in proportions 255
 - of ratios, 170
 - of regression coefficients using estimated weights 438
 - of transformed correlations, 187
- Welch-Aspin test, 115
- Wilcoxon signed rank test, 128
- Z or z , standard normal variate 51
 - z -transformation of a correlation coefficient, 185
 - tables, 558-559

